Rapid Deduction of Stress Response Pathways in Metal/Radionuclide Reducing Bacteria

Volume I
Project Overview and Research Plan

Director: Adam Arkin
Physical Biosciences Division
Lawrence Berkeley National Laboratory
Department of Chemistry
Department of Bioengineering
University of California, Berkeley
Howard Hughes Medical Institute

Co-Director: Terry Hazen
Earth Sciences Division
Lawrence Berkeley National Laboratory

Co-Investigators:
Alex Beliaeve, *Oak Ridge National Laboratory*
Carolyn Bertozzi, *University of California, Berkeley; Lawrence Berkeley National Laboratory, Howard Hughes Medical Institute*
Inna Dubchak, *National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory*
Matthew Fields, *Oak Ridge National Laboratory*
Hoi-Ying Holman, *Lawrence Berkeley National Laboratory*
Jay Keasling, *Lawrence Berkeley National Laboratory; University of California, Berkeley*
Martin Keller*, Diversa, Inc.*
Frank Olken, *National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory*
Anup Singh, *Sandia National Laboratory*
David Stahl, *University of Washington*
Dorthea Thompson, *Oak Ridge National Laboratory*
Judy Wall, *University of Missouri, Columbia*
Jizhong Zhou, *Oak Ridge National Laboratory*

April 2002

1 OVERVIEW

This proposal is in response to the Office of Science's Genomes to Life Program Announcement, LAB 02-13. It particularly addresses goals 2 and 4 of the announcement, which involve the characterization of regulatory networks in microorganisms, and the creation of data-driven, validated mathematical models of stress response to conditions commonly found in U.S. Department of Energy (DOE) metal and radionuclide contaminated sites. This proposal offers an integrated program of applied environmental microbiology, functional genomic measurement, and computational analysis and modeling that will seek to understand the basic biology involved in a microorganism's ability to survive in the relevant contaminated environments while reducing metals and radionuclides. The research and resources described in this proposal comprise the founding half of the Virtual Institute for Microbial Stress and Survival. A second project is being developed to exploit the studies elucidated here for biological threat reduction.

Our main focus is the microorganism *Desulfovibrio vulgaris* because of its metabolic versatility, its ability to reduce metals of interest to DOE, and its relatively easy culturability and molecular biology. However, because achieving our programmatic goals requires a comparative analysis of regulation among multiple bacteria in the environment, we will also study *Shewanella oneidensis* and *Geobacter metallireducens,* which follow different lifestyles than *Desulfovibrio.* Because a strong research community is already studying these microbes' behavior under the auspices of DOE's Microbial Cell program, we will coordinate with those teams to jumpstart the initial research of this program.

In the following overview, we outline the problem to be solved and discuss the challenges and management plan. The rest of this volume describes the operation and goals of different working groups (Core Teams) spanning applied environmental microbiology (Section 2), functional genomic measurement methodologies (Section 3), and computational/modeling efforts (Section 4). Finally, in Volume II of this proposal, we present budget information, biographical sketches, description of current facilities and resources, letters of collaborative support investigator support, and institutional forms.

1.1 Statement of the Problem and Mission Relevance

Metal and radionuclide contamination of soil and groundwater at DOE sites continues to be the major cleanup mission of the Department of Energy.

> When I became Energy Secretary—a little more than a year ago today—I was presented with the old plan for cleaning up our sites, which called for a timetable of some 70 years to complete and at a cost of $300 billion. That is not good enough for me, and I doubt it is good enough for anyone who lives near these sites. — *Spencer Abraham, Secretary of Energy 2/4/2002*

The Department of Energy currently has more than 350 cleanup projects, with a total life-cycle cost of $220 billion and a completion schedule of more than 70 years. Without major technical breakthroughs, the cost is expected to rise to $300 billion, an increase of over 36%, and could go much higher, according to DOE's Environmental Management Top-to-Bottom Review (2001).

The Department of Energy's cleanup of our country's nuclear production legacy represents the largest waste management and environmental restoration program ever undertaken. By conservative estimates, DOE has three million cubic meters of buried radioactive and hazardous waste and 75 million cubic meters of contaminated soil. DOE also has 475 billion gallons of contaminated groundwater. In addition, DOE has more than 20,000 nuclear weapons production facilities contaminated with radioactive materials, hazardous chemicals, asbestos, and lead (Figure 1.1) Some of these contaminated sites or facilities are no more than a few feet in diameter and less than a foot deep, still others are more then 20 square miles with a depth of more than 1,000 ft. below the ground surface (*DOE EQ Portfolio*, 2000; *DOE Paths to Closure*, 1999). Of this legacy waste, metals and radionuclides are the dominant cleanup problem and are found throughout DOE groundwater, soils, and sediments. More than 50% of DOE's facilities and 35% of its waste sites have radionuclide and metal contamination. In soils and sediments, radionuclides and metals are the highest-frequency classes of contamination by waste site, and more than 60% of

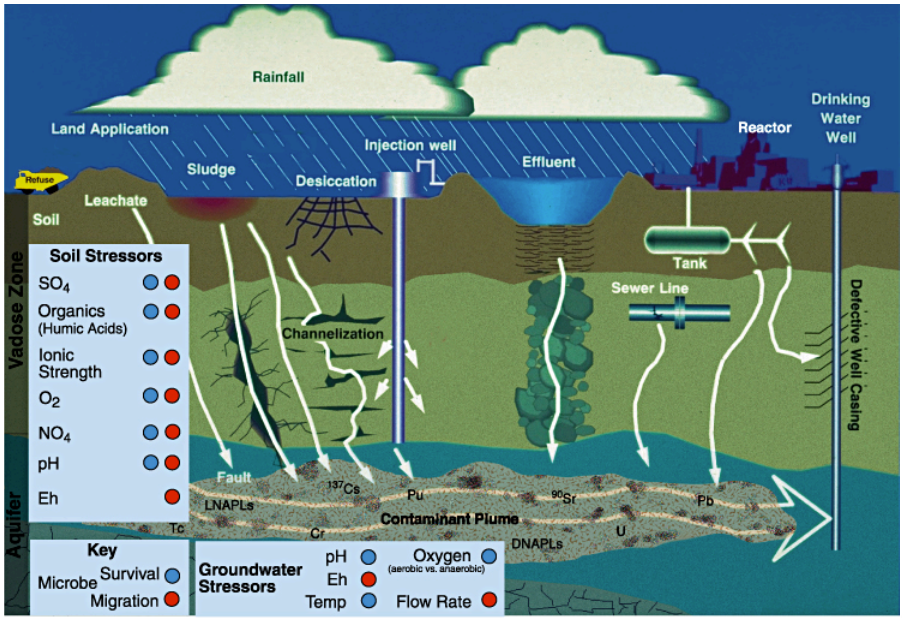


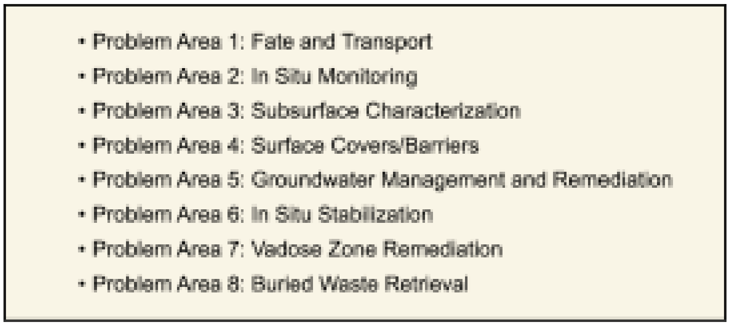
Figure 1.1. Sample contaminant plume consisting of mixed waste resulting from percolation from leaky tanks, landfills, basins, and trenches, as well as being formed through direct injection.

the groundwater at DOE facilities is contaminated with these types of waste. Metals and radionuclides in groundwater also are the highest-frequency compound class by waste site, with more than 50% having these contaminants.

Metals and radionuclides in the environment present the most difficult remediation problems because they cannot be destroyed, are inorganic, are reactive with soil and sediment constituents, and can remain hazardous at extremely low concentrations for centuries or indefinitely. Environments contaminated with metals and radionuclides present unique problems and have the highest remediation cost. The need for basic research that focuses on metals and radionuclides is further underscored by the recognition that radionuclides are a problem unique to DOE. And, unfortunately, because nuclear production was carried out by the DOE at DOE sites, it has not received the research attention or funding by other government agencies that solvents, fuels, and a few of the metal contaminants have received.

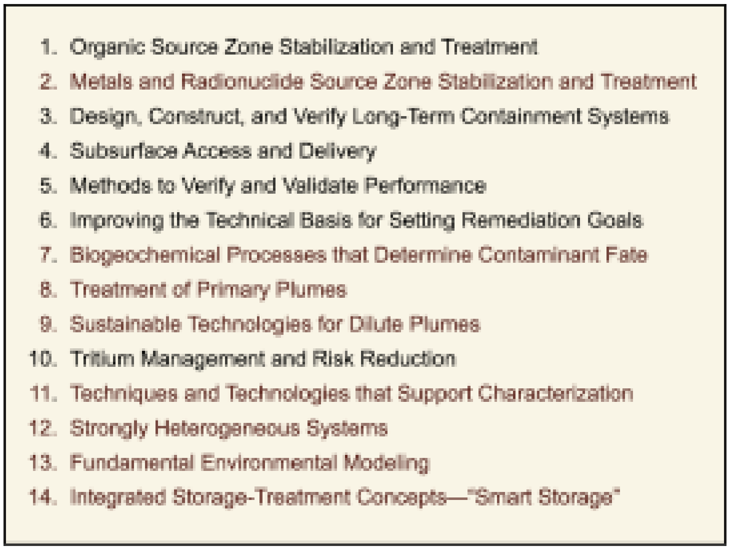
R&D Needs. Natural and accelerated bioremediation of radionuclides and metals has received the least research attention of any natural or accelerated in situ subsurface remediation process, yet this field holds the greatest promise for significant breakthroughs. A thorough understanding of subsurface mobilization and immobilization of radionuclides and metals will allow us to manipulate, stabilize, and predict long-term stability of these contaminants and their relative risk. This research will not only facilitate our overall understanding of our environment, but also will save DOE millions if not billions of dollars in cleanup costs. Despite the overall paucity of research in subsurface metal and radionuclide remediation and natural attenuation, DOE already has recognized the need to clean up this Cold War legacy waste and several years ago, through its Office of Biological and Environmental Research, established the Natural and Accelerated Bioremediation Research (NABIR) program.

What DOE's environmental restoration program still needs, however, is research and development in the areas of biogeochemistry and fundamental environmental process modeling of metal and radionuclide contaminant plumes (in soil and groundwater). Recently, Bechtel, the prime contractor for environmental restoration at five of the largest DOE sites, using a diverse team of technical experts, came up with a list of eight problem areas that need more R&D (Figure 1.2). This list includes better fate and transport models, better in situ groundwater and sediment remediation, and new methods for in situ stabilization of metals and radionuclides. At nearly the same time, the Subsurface Contaminant Focus Area in the Environmental Management's Office of Science and Technology put together another technical review team that came up with 14 remarkably similar R& D technical targets (Figure 1.3). These targets include new technologies for metal and radionuclide source-zone stabilization, fundamental environmental modeling, and biogeochemical processes that determine contaminant fate.



- Problem Area 1: Fate and Transport
- Problem Area 2: In Situ Monitoring
- Problem Area 3: Subsurface Characterization
- Problem Area 4: Surface Covers/Barriers
- Problem Area 5: Groundwater Management and Remediation
- Problem Area 6: In Situ Stabilization
- Problem Area 7: Vadose Zone Remediation
- Problem Area 8: Buried Waste Retrieval

Figure 1.2. Bechtel's Technology Panel Results.



1. Organic Source Zone Stabilization and Treatment
2. Metals and Radionuclide Source Zone Stabilization and Treatment
3. Design, Construct, and Verify Long-Term Containment Systems
4. Subsurface Access and Delivery
5. Methods to Verify and Validate Performance
6. Improving the Technical Basis for Setting Remediation Goals
7. Biogeochemical Processes that Determine Contaminant Fate
8. Treatment of Primary Plumes
9. Sustainable Technologies for Dilute Plumes
10. Tritium Management and Risk Reduction
11. Techniques and Technologies that Support Characterization
12. Strongly Heterogeneous Systems
13. Fundamental Environmental Modeling
14. Integrated Storage-Treatment Concepts—"Smart Storage"

Figure 1.3. SCFA Technical Targets.

The effective implementation of remediation strategies and the use of natural attenuation for the cleanup of DOE sites depend on understanding critical chemical, physical, and biological processes. Given that many DOE sites are in remote areas, with few risk receptors and many of these plumes appearing to be stationary or naturally attenuating, there is a dire need for the enabling science to prove that natural attenuation and in situ bioremediation are safe and practical technologies for these sites. These remediation strategies, as well as a better understanding of the environmental risk that metals and radionuclides pose in these situations, will allow DOE to decrease risk to the public, thereby encouraging public trust, decreasing closure schedules by decades, and saving billions of dollars in cleanup costs.

Our proposed Genomes to Life project directly addresses the need to provide biological solutions for DOE missions. The GTL roadmap specifies the need "to provide knowledge about using natural populations of microorganisms to degrade or immobilize contaminants and accelerate the development of new, less costly strategies for cleaning up DOE waste sites." Our project has also been designed to address each of the three aims in Goal 2 of the GTL roadmap: (1) characterize gene regulatory networks, in terms of mapping microbial regulatory circuitries; (2) connect regulatory network properties with their biological outputs; and (3) develop computational modeling tools to predict the dynamic behavior of natural or designed regulatory networks.

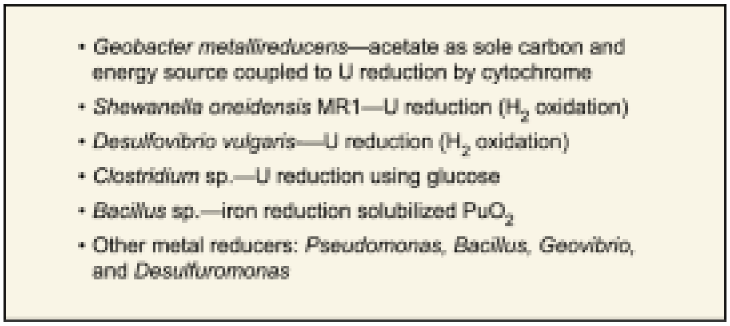
Collateral Research Needs. In 2001, the American Academy of Microbiology made several related basic research recommendations in their "Geobiology: Exploring the Interface Between the Biosphere and the Geosphere," a colloquium co-sponsored by DOE. In their report they recommend: (1) support laboratory studies of basic biological sciences, especially microbial diversity, physiology, and genomics, which represent the basic science backbone of the field; (2) improve methods for defining and detecting chemical signatures and other biomarkers that indicate the presence of past and present life; (3) encourage further exploration of how human beings might exploit the geobiological capabilities of microbes for society's benefit, such as in cleaning up polluted areas or extracting valuable resources from the environment; and (4) promote studies of geobiology that will enable the finding of extraterrestrial life.

The survival of pathogens (e.g., *Bacillus anthracis* spores) in both the environment and host will also require similar stress regulatory pathways to those that allow metal and radionuclide reducing bacteria to survive in the subsurface and in the extreme environments encountered at many DOE waste sites (oxidative stress, for example is one of the main antimicrobial strategies employed by the immune system). Pathogen survival in both the host and the environment is such an important science need for better handling of bioterrorist events, emerging diseases, and treating human disease that we are developing a parallel and complementary program to take advantage of the stress pathways elucidated by the proposed project.

The Focus. This program addresses DOE cleanup and collateral research needs. Our focus is on experimentally elucidating and computationally modeling the stress response pathways of three target metal and radionuclide reducing organisms: *Desulfovibrio vulgaris, Shewanella oneidensis* MR-1, and *Geobacter metallireducens*. Microbial metal reduction plays an important role in biogeochemical cycling of carbon and nitrogen, as well as in the bioremediation of metals, radionuclides, and organic contaminants. A number of bacteria have demonstrated the capability to reduce radionuclides and other metals (Figure 1.4). Numerous microorganisms capable of coupling the oxidation of organic compounds to the reduction of metals have been isolated and studied from the standpoints of physiology, ecology, and phylogeny.

However, the success of various bioremediation approaches largely depends on our understanding of regulatory mechanisms and cellular responses to different environmental factors affecting the metal reduction activity in situ. Microorganisms are often exposed to multiple stress conditions in situ, and the pleotrophic effects on community structure and functional gene pools are most likely mediated via the stress response systems of individual microorganisms. Our recent results at Oak Ridge National Laboratory suggest that definitive relationships between geochemical parameters and the microbial communities are difficult to elucidate**.** When the distribution of cloned nitrite reductase genes, *nir*K and *nir*S, was compared to geochemical measurements (levels of nitrate, uranium, pH, heavy metals, and organic carbon) at six groundwater sites, the cloned genes appeared to cluster in relation to nitrate levels. However, the types of genes (sequence diversity) could not be predicted with high confidence when a linear or logistic model was used. An artificial neural network improved the prediction ability of sequence distribution, and this result indicated that the relationships between microorganisms with a *nir*K or *nir*S gene and site geochemistry were nonlinear (Table 2.1). In addition, diversity indices for the gene pools at acidic, contaminated sites were sometimes higher than the background site, and the same predominant sequence was observed at the acidic, contaminated sites as well as the background site. These results indicate that the effects of multiple stresses on microbial communities and functional capacities are complicated and not well understood.

Desulfovibrio vulgaris, Shewanella oneidensis, and *Geobacter metallireducens* represent three different groups of organisms capable of metal and radionuclide reduction whose complete genome sequences were (or are being) determined under the support of DOE-funded projects. Utilizing the available genome sequence information, we will focus our efforts on studying the stress response pathways in these microorganisms, which are induced by various environmental factors such as oxygen, temperature, and nutrient concentrations. (Note: Our primary focus



- *Geobacter metallireducens*—acetate as sole carbon and energy source coupled to U reduction by cytochrome
- *Shewanella oneidensis* MR1—U reduction (H_2 oxidation)
- *Desulfovibrio vulgaris*—U reduction (H_2 oxidation)
- *Clostridium* sp.—U reduction using glucose
- *Bacillus* sp.—iron reduction solubilized PuO_2
- Other metal reducers: *Pseudomonas, Bacillus, Geovibrio,* and *Desulfuromonas*

Figure 1.4. A list of metal-reducing bacteria (Banaszak, Rittman, and Reed, 1999).

will be on *D. vulgaris*, since much less is known about this organism. However, our initial work also will involve *S. oneidensis* and *G. metallireducens* because of the large amounts of data already available on these two bacteria. Stress could have significant effects on Fe(III) reduction in *Shewanella oneidensis*, a major pathway for biotransformation of metals and radionuclides in the environment. A number of studies have demonstrated that Fe(III) reduction by *Shewanella* is linked to an electron transport chain involving cytochromes and other electron carriers. [*Geobacter* has also been shown to completely oxidize fermentation products to carbon dioxide under Fe(III) respiration. Terminal electron accepting processes (TEAPs) for *Geobacter* and other bacteria are known to switch as the environment changes. Although all three of these bacteria are Proteobacteria, they differ in that *Shewanella* is a microaerophile, while the other two are anaerobes. In addition, *Desulfovibrio* is common in eutrophic environments, while the other two are more common in oligotrophic environments. The similarities and differences among these organisms should allow for a better comparison and delineation of stress regulatory pathways and conserved components that are shared across species and habitat boundaries.

The overarching goal of the proposed research is to develop criteria for monitoring the integrity (health) and altering the trajectory of an environmental biological system (process control). To achieve this requires a more complete understanding of how the biological "units" comprising the system are organized, regulated, and linked in time and space (genes, genomes, cells, populations, communities, and ultimately, ecosystems). Key to these objectives is a more complete understanding of stress response systems and their environmental context.

1.2 Overview of Approach

The goals for this project fall into two categories: applied and pure. The applied goals are: to develop better methods (1) for determining the ability of natural soil microbes to attenuate and immobilize metal and radionuclide contamination and (2) for applying biostimulatory agents to maximize these effects. The pure goals are to understand, from a physical-chemical, control-theoretical, and evolutionary point of view, the structure, function, and dynamics of the pathways involved in the biogeochemistry of soil microbes under a wide variety of conditions. In this way, we will ultimately be able to predict, control, and design pathways in these organisms for specific decontamination goals or provide an enabling knowledge to predict the efficacy of natural attenuation processes. Fulfilling the applied goals rests in large part on the successful completion of the pure goals. However, the two are optimally integrated in an iterative loop, each cycle of which yields better methods and models. (See Subsection 1.2.2 for a detailed description of this cycle.)

Because of the scarce data on the target organisms, we must begin this project by pursuing the pure goals. First, we will measure in physical detail the time-dependent activity of as many pathway components as possible under a variety of conditions, stresses, and other perturbations. From both perturbation-response data and direct measurements of molecular interaction, we will then deduce pathways involved in the stress response of target organisms during the task of metal reduction. The same perturbation-response data are also a necessary precursor for understanding how changing soil conditions and the applications of external stimulatory agents to these organisms will change and control their behaviors. The following subsections explain this choice of experimental design and why we need perturbation data from more than one organism. We also provide a roadmap describing the challenges we face in achieving this project's goals.

1.2.1 Why Many Pathways from Multiple Organisms in Great Detail?

By creating detailed causal/physical models of the stress response pathways, we will learn what the principles of control in these pathways are at a molecular level. In order to do this we must have extensive measurements of the time-dependent changes in activity for all molecular players and their interactions. It does not suffice to have only the cis-regulatory structure that comprises the immediate control of gene expression, nor only the protein-protein interactions. Ideally, all the state-dependent interactions among the cellular components should be traced.

This is because bacterial regulatory networks are less stratified than it might seem at first. For example, the sporulation initiation pathway in *Bacillus subtilis* (a cryptobiotic pathway) shown in Figure 1.5 demonstrates that all types of interactions can play a major role in cellular decision making. That is, control of gene expression, signal transduction, secretion of small peptides and proteins, and metabolite consumption are all coupled to arrive at the final decision to sporulate. The network in Figure 1.5 was experimentally discovered over a 25-year period and contains a large amount of both genetics and biochemistry. Even so, the exact control of sporulation has still not been solved. To match and exceed this level of detail in our target organisms will require a concerted effort to measure as many cellular species as possible, as well as their interactions and their behaviors, under a wide variety of external conditions and mutant states. The functional genomics bioanalytic pipeline described below is designed to address this need.

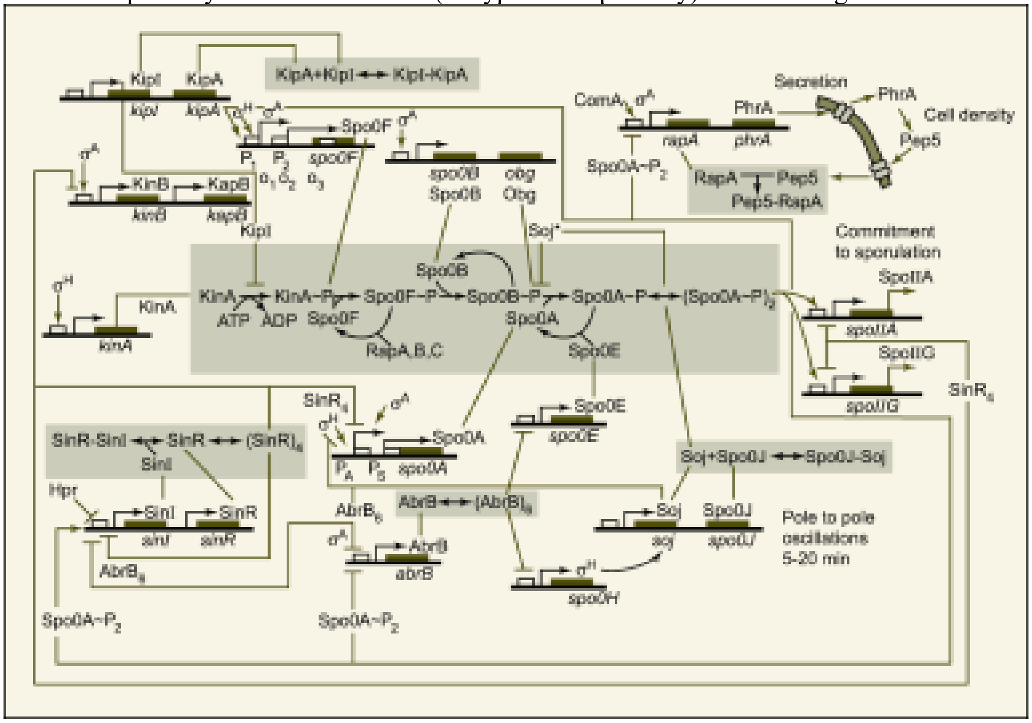


Figure 1.5. The sporulation initiation pathway in *Bacillus subtilis.*

The sporulation initiation pathway diagram in Figure 1.5 is artificially truncated. Many other molecular species are integrated into the function of this pathway. In addition, many of these other species, for example ComA, are involved in other pathways. Figure 1.6 shows how the seemingly innocuous notation for ComA in Figure 1.5 (middle, upper right of that figure) can lead into more complex networks. One of the central challenges of network biology will be development of methods that discover modular structure in these pathways, if such exists. Certainly the pathways can at least be conceptually broken up by overall function. For example, the stress response pathways in *Bacillus subtilis* interact with each other in such a way that a perturbed population of cells splinters into a number of different cellular behaviors, each cell implementing one or more stress responses (Figure 1.7). These stress responses modulate each other so that incompatible behaviors such as sporulation and chemotaxis are not expressed simultaneously (or contemporaneously). This sort of interaction among pathways is more than simple cross-talk. Cross-talk, at least in engineering, implies parasitic signals that move uncontrolled among closely spaced systems. Yet in this discipline, cross-talk has strong biological implications. This strong coupling and the resultant population heterogeneity necessitate the study of more than one stress response and the development of experimental methods to follow which cells choose which responses and, then, to deconvolve these responses from population-based data. Indeed, a central goal of the theoretical work is to derive formal methods for determining modularity in these networks so that subsystems can be tested and modeled without complete information about the “cross-talked” systems.

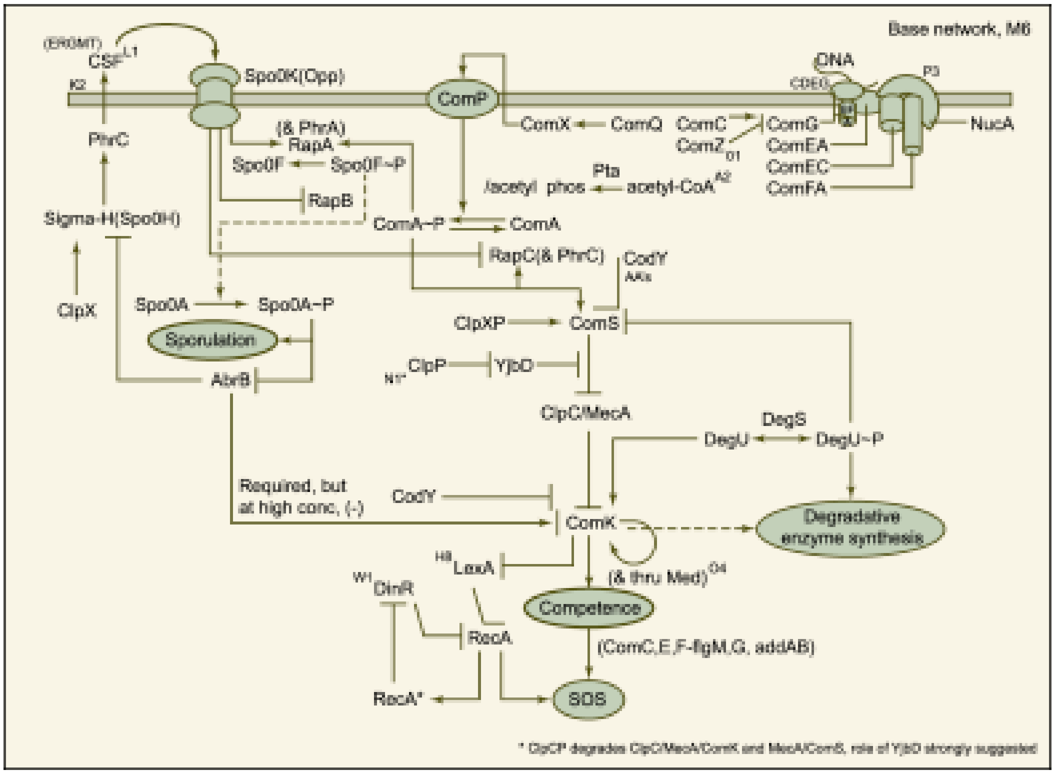


Figure 1.6. Coupling of sporulation to other pathways.

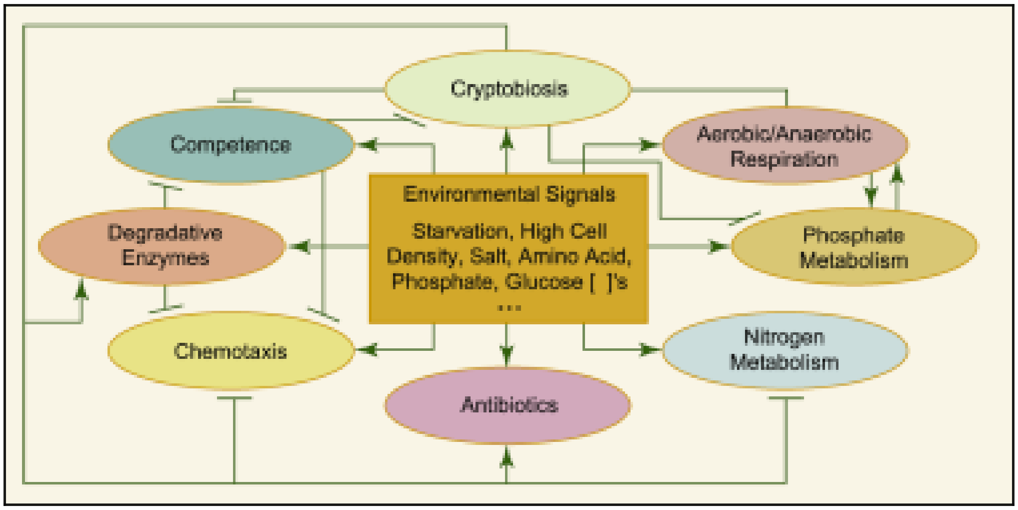


Figure 1.7. Different stress responses caused by splintering of perturbed population of cells after interaction of stress response pathways in *Bacillus subtilis.*

Figure 1.7 shows the interaction among stress response pathways within a particular organism (in this case *B. subtilis*). However, although other organisms may have homologous pathways to *B. subtilis*, the regulation among and within these pathways is likely to be different. As an example, in a recent comparative study of the chemotactic pathways in *Escherichia coli* and *B. subtilis*, it was found that, although the major complement of proteins was the same between the two organisms, there were both component differences and important differences in regulation.

Figure 1.8 shows a schematic diagram of those differences. The study concludes that the systems respond very differently to perturbation in their machinery. Two classes of question arise in the face of this information. First, how do the differences in perturbation response behavior lead to differences in the endogenous ability to track chemical signals, and what do these dissimilarities imply about the differences in control that may be achieved by external (unnatural) signals? Second, are these differences traceable to the particular environments in which these two organisms live? *E. coli*, a gram-negative microbe, is an enteric organism that needs to live in animal hosts as well as in external environments, whereas *B. subtilis,* a gram-positive, spends most of its life in soil. Are there evolutionary roots to the disparate regulatory structures, or are they merely artifacts of evolutionary drift? In the case of bioremediation applications, it is especially important to know about and understand these differences and how they lead to controllability of a behavior by which many different microbes, all with different instantiations of these stress response pathways, are either competing and cooperating within the same niche or, of necessity, operating in different niche environments.

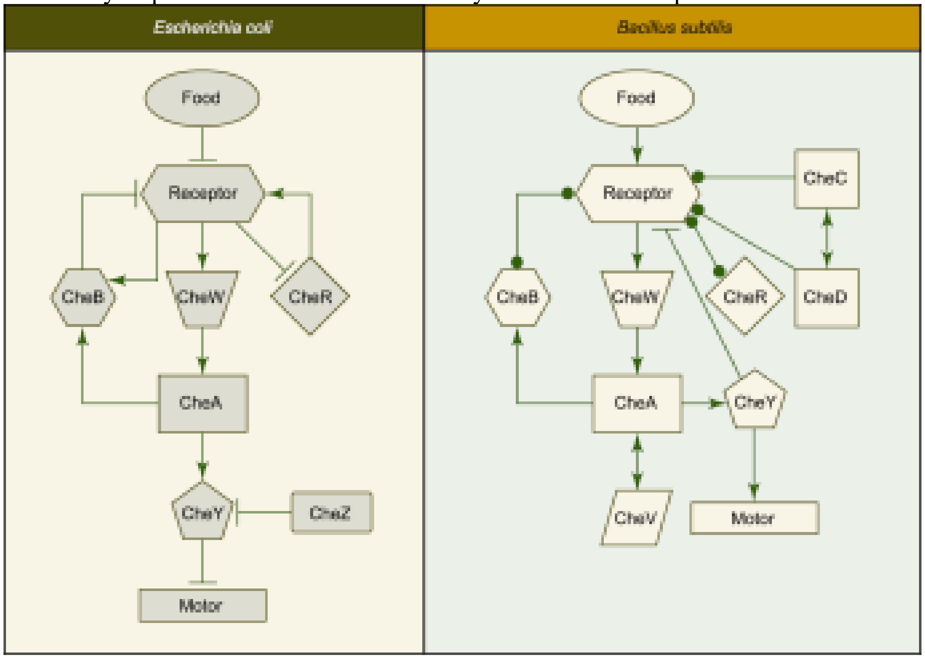


Figure 1.8. Comparison of the schematics for chemotaxis in two organisms.

One of the main hypotheses of the proposed work is that it is important to understand differential regulation among the microbial community in the contaminated sites if we are to understand the course of natural and bio-stimulatory remediation. During the course of exposure to contaminants or to biostimulatory strategies, the different microbes, with their different regulatory strategies, will respond differentially, and each may have different times and conditions of dominance for metal and radionuclide reduction. Any designed control should take the changing landscape of organisms into account. These changes also have profound implications for the overarching biogeochemistry of the subsurface in general.

Thus, although these three key organisms will provide core models of the stress response pathways, obtaining information on the regulatory strategies of other organisms and their relative overall responses to different conditions will provide further confidence in the estimation of regulatory structure differences across and between niches and of how these regulatory differences lead to more or less efficiency in decontamination and alteration of a variety of biogeochemical processes. We propose to examine the diversity of regulatory strategies, both by comparing our three key organisms and by identifying the community of other microbes in the environment. Large-insert cloning of pathways from a number of these organisms, homologous to the pathways identified in our key microbes, will lead to identification of both conserved and divergent sets of proteins and DNA cis-regulatory elements. Deducing which pieces of a given pathway are central to a given function can be done by observing which factors are conserved across niches. Within-niche variation of regulation should give insight both into the plausible flexibility of the regulatory network in achieving similar goals and the differential regulation of organisms competing for resources within the same niche. Conserved differences between niches should yield insight into niche-specific regulatory strategies. Combining this information with experimental measurements on the population dynamics of the community under different stress conditions will yield a mapping between regulatory strategy and behavioral phenotype. Models of the different strategies, modified from the highly tested models of the target organisms, will yield hypotheses for the role of each regulatory difference in the survival of the microbe and the fabric of microbial community interaction with the environment that controls biogeochemical processes.

1.2.2 The Research Cycle

The program's fundamental research cycle is shown in Figure 1.9. The starting point in this diagram is the collection of soil samples (box 1, upper left). These samples will be collected from a number of different DOE waste sites, including those at LLNL, and the NABIR Field Research Center at Oak Ridge (our primary source). These soil samples will be analyzed first for the content of nutrients, metals, radionuclides, pH, and other important soil states and stressors (box 2). Such chemical analyses will be done both before and during treatment or under induced-stressed conditions. From the analysis of these soil samples, a set of experimental soil simulators will be built and tested at a variety of scales (10 µm to 1 m) in a variety of configurations (static, chemostat, soil column, batch reactor). These are actual laboratory culture instruments in which bacterial growth may be observed in controlled conditions with a more or less defensible relationship to soil conditions at the contaminated site. Bacteria grown in these simulators (either seeded from isolated target organisms or from natural raw soil samples) can then be harvested during the course of treatment or control experiments. These simulators will also be used to explore by direct examination (Synchrotron FTIR, phospholipid fatty acids, fluorescent in situ hybridization using flow cytometry) changes in stress responses that affect population, microbial community, and biogeochemical interactions. Examples of stressors include the metals and radionuclides themselves as well as phosphates, nitrates, low or high carbon and oxygen, pH, salt, and numerous other determinants of the cell's ability to process the contaminants. Thus, boxes 1–4 fall under the purview of the Applied Environmental Microbiology Core. Box 4, however, is shared with the Functional Genomics Core.

Task 2 is to directly measure the interaction among the biomolecules in the system. The two primary technologies we will use are phage-display interaction trapping and protein-crosslink mass spectrometry. These are complementary methods for detecting such interactions involving two or more molecules. They are used, in concert with the Task 1 data, to propose a regulatory network structure.

Task 3 involves probing interactions predicted to be critical after analysis of data from Tasks 1 and 2. In this task, synthetic chemical inhibitors (or activators) for this interaction are developed by the combinatorial chemistry facility (box 12, Figure 1.9). The inhibitors are then used singly and in combination to dissect precisely how, in wild-type cells, perturbations of these key interactions could lead to synergetic effects on the ability of the organisms to process contaminants. The system identification task will be aided by these results, and the chemical toolkit that is developed may actually prove useful in the field for biostimulatory treatments.

Task 4 is really the central phenotypic measurement pipeline in that the goal is to measure how different populations of microbes, including the targets, grow and die under the different perturbation conditions (box 7, Figure 1.9). The genomic arrays are designed to track how many of each identified microbe are present in the culture sample. Complementary information will be derived from the flow cytometry and phospholipid tracking measurements made in the Applied Environmental Microbiology Core. These, in combination with the Task 1 and 2 data, provide an experimental mapping between the molecular activities and the overall ability of the organism to survive and operate. The Computational Core will develop both statistical and causal/physical models to quantitate and explain this mapping (see below).

Task 5, together with the data from Task 4, is geared to understanding how different regulatory strategies are employed to survive in different niches and compete within niches. This task involves using predictions of which proteins and pathways are involved in the cellular stress response and decontamination response and then cloning out homologous pathways from other microbes in the environments surrounding the targets. The large-insert clones will contain about 50 kb of material surrounding the target genes. Thus, both the other proteins in the operon and the intergenic regions in which the regulatory elements reside will often be obtained. The Computational Core will use these fragments to predict the different regulatory strategies used by different organisms in coordinating their stress response and to predict how these might be related to the measurements in Task 4. Together, these will allow construction of the microbial population models that will allow exploration of conceptual models for natural attenuation, bioremediation, and the fundamental biogeochemistry of the test environments.

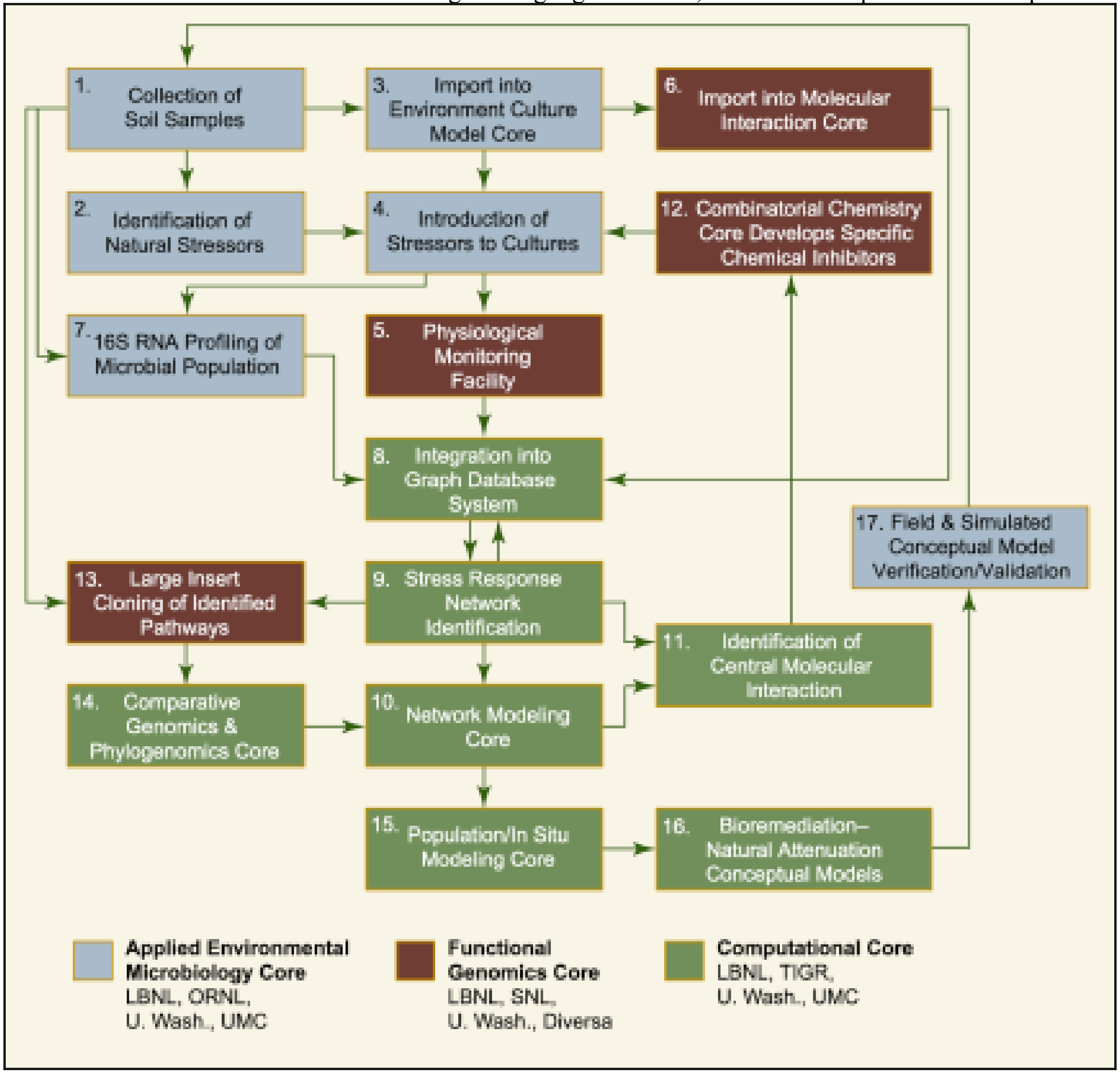

1. Collection of Soil Samples
2. Identification of Natural Stressors
3. Import into Environment Culture Model Core
4. Introduction of Stressors to Cultures
5. Physiological Monitoring Facility
6. Import into Molecular Interaction Core
7. 16S RNA Profiling of Microbial Population
8. Integration into Graph Database System
9. Stress Response Network Identification
10. Network Modeling Core
11. Identification of Central Molecular Interaction
12. Combinatorial Chemistry Core Develops Specific Chemical Inhibitors
13. Large Insert Cloning of Identified Pathways
14. Comparative Genomics & Phylogenomics Core
15. Population/In Situ Modeling Core
16. Bioremediation–Natural Attenuation Conceptual Models
17. Field & Simulated Conceptual Model Verification/Validation
Applied Environmental Microbiology Core
LBNL, ORNL, U. Wash., UMC
Functional Genomics Core
LBNL, SNL, U. Wash., Diversa
Computational Core
LBNL, TIGR, U. Wash., UMC

Figure 1.9. A flow diagram of the research cycle. The coloring of the boxes indicates the subgroup to which each task is assigned

The Computational Core has a number of diverse tasks, ranging from setting up quality control in the Functional Genomics Core such that the data are useful for modeling, to designing the data storage and query infrastructure, and making models of the pathways. The majority of the boxes in Figure 1.9 are tasks in this core (boxes 8–11 and 14–17). Because of the diversity of the data generated by the Functional Genomics Core and the goal to relate these to pathway structure and models, a novel database technology will be developed based on the graphical nature of this information (box 8). Specialized data entry, query, and curation tools will be built so that the Functional Genomics Core and Applied Environmental Microbiology Core can easily interact with this information. The tasks and policies developed as part of the box 8 activities comprise the data and information management plans. See Section 4.3 for a full exposition.

However, before such a database can be useful, the data it is populated with must be reliable. Thus, in collaboration with the Functional Genomics Core, quality control protocols will be developed so that only reproducible, well-scored information is entered into the database. Once accomplished, methods for identifying molecules whose state is significantly changed in a perturbed condition compared to control conditions will be designed. From these and other data, initial statistical and causal models of the regulatory network structure will be made and scored against data (boxes 9 and 10).

Once a network structure has been proposed, chemical and physical principles, together with directed measurements from the Functional Genomics Core, can be used to produce and validate dynamical models of pathway function (box 10). These lead to predictions of how the organism survives, grows, and operates under different conditions that will be compared to the results from box 7 (Task 4 above).

Once well-understood dynamic models of the key organisms are produced, predictions of cis-regulatory and protein-complement differences among the surrounding microbial community from comparative sequence analysis (box 14) will be entered into the models to see if these are enough to explain the growth differences observed in Task 4. If not, the rest of the model will be modified to produce hypotheses for nongenetic regulatory differences among the pathway elements sufficient to explain the Task 4 results. From these, combined models of the population behaviors of the measured community can be made (boxes 10 and 15).

The population models will then be used as a substrate for testing different ideas for bioremediation strategies. In one mode, automatic tools will explore the response surface that characterizes how a model observable (such as the total amount of metal reduced over a set period of time) depends on all the parameters of the model. This will lead to suggestions about which interactions in the network will be most important to target if a biostimulatory strategy is to be used. In another mode, the Applied Environmental Microbiology Core can use the models to test various existing and proposed bioremediation strategies and quantitate differences between predicted natural attenuation and what derives from biostimulation. The strategies deemed best during this simulation process would be implemented in the laboratory and the field to see if the model and reality compare favorably (boxes 15–17). If not, this is a further piece of data that re-enters the analysis from the beginning. We expect there to be a number of such experimental/computational cycles.

This basic research plan thus relies on a very tight integration of the three teams and their expertise and technologies. Therefore, a careful assessment of the scientific and computational challenges to carrying out each stage of this research and a good management plan are necessary. Subsections 1.3 and 1.4, address these issues.

1.3 Scientific and Computational Challenges

The individual Core Team sections (Sections 2–4) will examine in detail the challenges of completing each task outlined above. However, a number of these are summarized in this overview to underscore some of the innovative research that must be accomplished to overcome the roadblocks.

1.3.1 Sample Collection and Simulation

Collection of subsurface samples is a major challenge. Because subsurface samples are collected by drilling techniques, all samples are mechanically disturbed and quite often also disturbed by heat and fluids associated with most drilling processes. Recovery, packaging, and transport can also severely alter the conditions of the environment by introducing oxygen or cutting off sources of nutrients for periods long enough to cause stress or alter the biogeochemistry. In addition, significant health and safety concerns must be addressed for the collectors at contaminated sites, and permits must be obtained for transport and handling. Fortunately, the DOE NABIR Field Research Center (FRC), which was established at Oak Ridge two years ago, has dealt with these issues in order to supply investigators with defensible samples. The NABIR FRC will be our primary initial source of environmental samples.

Creating defensible culture conditions, i.e., simulating those conditions that mimic the environments where these stress responses are important will also be difficult. Control of redox conditions, pH, and allochthonus sources of microbes is daunting under the best conditions. In these experiments, we will have to control these conditions and at the same time try to maintain stress states that limit most microbial growth. The Functional Genomics Core team's need for large amounts of biomass will necessitate large numbers of soil columns and chemostats and meticulous quality control and quality assurance. Indeed, in some cases we may need to mimic environments that are stressed to the point that nonculturable microbes are dominant.

Because the accuracy and direction of the Functional Genomics Core studies and the Computational Core depend on these environmental simulators, quality control and assurance will require robust, rapid, and direct methods for identification of target organisms, communities, and even viable but nonculturable organisms. We plan to use a variety of techniques (e.g., microarrays, FISH flow cytometry, PLFA, Synchrotron FTIR) to meet this challenge.

1.3.2 Measurement Pipeline

Quality and consistency of data are the primary challenges faced by the Functional Genomics Core. It is well known that significant changes in gene expression are necessary for detection using DNA array technologies. These same problems occur in the other high-throughput components of this proposal: proteomics and metabolomics. However, both of these technologies suffer from a lack of maturity and application that DNA array technologies have overcome. Just as with DNA array technologies, all measurements using these other technologies will be taken relative to a set of controls; thus, only relative changes will be monitored. Fortunately, stress responses are defined relative to a "normal state," so these relative measurements should be ideal. In all of these technologies, we have chosen investigators with significant experience in each of these areas. These investigators recognize these data quality issues and have addressed them in the relevant areas of this proposal.

Another significant hurdle will be standardization of data. The distinguishing feature of transcriptomics, proteomics, and metabolomics is the large amount of data that can be collected. Because each of these techniques will be performed by different investigators (experts) in different locations, it could be difficult to compare data from transcriptomics, proteomics, and metabolomics experiments. To ensure that data from these technologies are comparable and consistent, experiments (cell culture, application of stress) will be performed in as identical a manner as possible. In some cases, cell-culture and stress-response application will be performed in one location, and samples will be sent to the various laboratories for transcriptomic, proteomic, and metabolomic analyses. In addition, all data will be annotated consistently among the various laboratories so that the information can be easily compared. As the data are collected and annotated, they will be posted on a Web site with limited access to the team members, so that team members may compare data from the various techniques.

A less significant hurdle will be management and analysis of high-throughput data collection. When all aspects of the proposed research are running at full speed, we envision that a significant number of samples must be processed. For example, several different stresses will be applied to each of the three bacteria, and each of these stresses will have a time-dependent response. Thus, many samples will need to be analyzed for their transcript, protein, and metabolite profiles. Fortunately, each of the techniques is eminently scalable with autosamplers and robotics, so that the largest challenge will be data management. The data management challenges will be addressed by the Computational Core.

Creation of mutants and interruption of stress response pathways also present challenges. The first and most significant challenge is the establishment of a genetic system for *Desulfovibrio*. Fortunately, we have recruited the help of an expert in *Desulfovibrio* genetics (Wall). Further, the other two organisms to which we will compare stress responses in *Desulfovibrio* have genetic systems that will allow introduction of mutations in stress response genes. The second significant challenge in stress response pathway interruption is finding chemical mutants that prevent elicitation of a particular stress response. For this purpose, we have chosen a high-throughput chemistry approach that will allow us to synthesize and test a large number of chemicals simultaneously.

1.3.3 Computational Core

The Computational Core has many challenges to overcome in its four major areas of operation: (1) database creation and management, (2) data quality control and analysis, (3) comparative functional genomics, and (4) model deduction and analysis.

Area One, Database Creation and Management. Many canonical areas will need to be addressed. First, there is the problem of how to define a database schema flexible enough to hold the diverse data types coming off the measurement pipeline and the design of the data interchange standards. This will be addressed first by adhering to as many existing standards as are available. Few standards, however, exist: MGED has produced standards for microarray information; SBML and CELLML have produced interchange standards for cell models (that unfortunately don't include good models for genetic information or relationship to experimental data); NCBI/NLM have produced an extensive relational schema that holds most of the standard data types, from microarray to interaction information, to sequence, to literature reference. It is a rather shallow database in some ways, however, and unwieldy when used for pathway information. It also has no facilities for modeling. We will fill in the gaps in data interchange where necessary and have already begun to modify the NCBI databases we can use to better handle linkages among the disparate data types and pathway information. However, certain desirable queries for model analysis and pathway comparison require special operations on the graph structure of the pathway. These queries are inefficient or impossible in the standard relational framework; thus, we will design and implement a novel graph database designed to be maximally efficient for such queries.

Second, there is the problem of data entry, query, and curation, with particular emphasis on the latter. In the curation process, data are normalized and quality controlled, and revisions are entered and resolved. The Computational Core proposes a set of curation policies based on assigned responsibility for different data types and their relationships to individuals who specialize in the relevant data type, and on an authority model and revision tracking system that allows changes to this information to be tracked only by qualified individuals.

Area Two, Data Quality Control and Analysis. It is imperative that the data derived from the Functional Genomics Core are reproducible and cross-comparable. However, microarray technologies are notoriously noisy, as are interaction traps and even separation/mass-spectrometry measurements. In Section 4 of this proposal, the Computational Core will propose a number of techniques for dealing with some of these issues and for attaching statistical confidence to each measurement.

Area Three, Comparative Functional Genomics. There is the central problem of determining true homology among proteins in the stress response pathways of multiple organisms and of identifying the short cis-regulatory sequences central to the stress response strategies. The homology problem is a long-standing issue, and the respective genome projects of the target organisms are largely dealing well with this issue. However, when faced with the analysis of the large insert clones, we will have to follow the annotation pipelines of these other projects as well as use other methods of our own design. For example, one of our collaborators, Jonathan Eisen at TIGR, has developed methods using phylogenetic comparison to better classify new proteins into functional categories. The detection of regulatory elements has been the subject of much recent research. Methods based on syntenic alignment, gene-expression-correlated intergenic region analysis, and other probabilistic classifiers all have moderately high success rates. These will be combined in this study to provide a more robust estimate of the existence of these regulatory regions. When incorporated into a model, their consistency with the molecular profiling and population growth data will be possible.

Area Four, Model Deduction and Analysis. The making of the models themselves presents a challenge. Depending on the degree of causal structure and physical description, more or less data are necessary to construct a model. Depending on the questions asked, the type of model to be used may range from a statistical/ phenomenological model to a highly detailed stochastic chemical kinetic model. The challenge will be to develop a framework that allows modeling at these different levels of abstraction parameterized by different sorts of data and to develop a principled scoring scheme that rates how well a model reproduces the input information and explains/ predicts other experiments. Section 4 describes the various approaches to this problem and the tools that will finally be produced based on validated models to aid in creating conceptual models of population growth, survival, and metal/radionuclide reduction.

1.3.4 Conceptual Models and Implementation

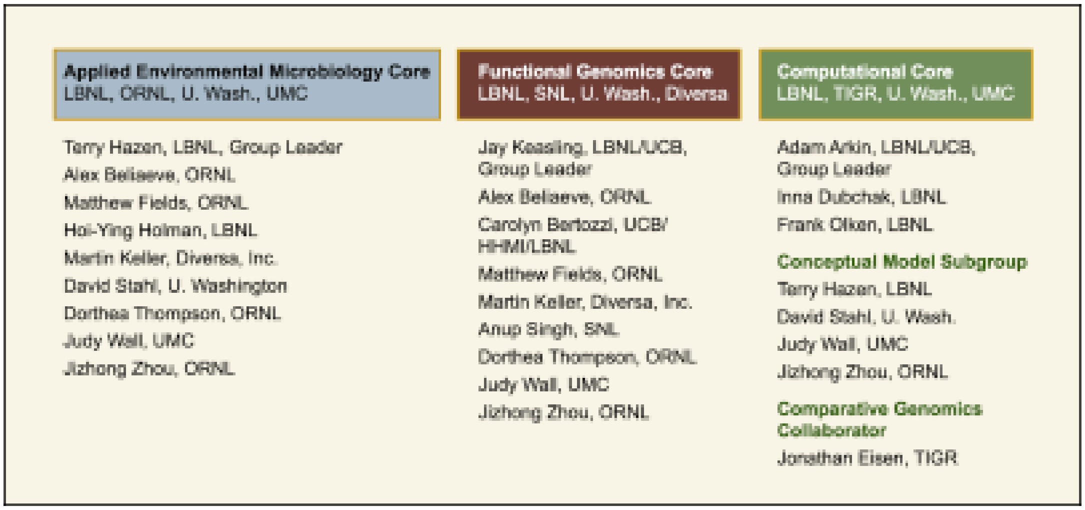
Conceptual models of subsurface environments are notoriously difficult to construct in a manner that conveys all of the major pathways and linkages between major components. Yet a good conceptual model of the subsurface environment is requisite for making good environmental stewardship decisions. A well-defined conceptual model is the primary building block of all numerical models for contaminant fate and transport predictions. The subsurface environment can be affected by infusion from various sources and controlled by the ambient types of rock and sediment, which can affect liquid and gas flow, geochemistry, nutrient supply, and other stressors. Indeed, episodic events such as rainfall and snow melt, subsurface special heterogeneity, and presence of electron donors for bacteria can cause dramatic changes in the overall redox conditions, and this would strongly influence subsurface reactive chemistry. We recently showed that diffusion-limited areas could increase microbial reduction of toxic Cr(VI) to nontoxic Cr(III) several times in soil aggregates in the presence of adequate carbon sources.

Testing of these conceptual models in both laboratory simulations and field tests is also a daunting task, given the challenges described for the culture and simulations, and the difficulty in finding gradient stressors or particular perturbations at contaminated sites that are not complicated by other contaminants or factors that mask the effect. Most DOE sites contaminated with radionuclides have low pH and high nitrate levels, but they often also have high concentrations of solvents such as PCE or carbon tetrachloride, which can be toxic or stressing by themselves. Little is known about what synergistic effects this toxic milieu could have on the biogeochemistry of a site. Only by carefully factoring out each of these parameters in a quantitative modeling context will we be able to show how our conceptual understanding of stress interaction at these sites can be used to implement the best stewardship of the site for human health and the environment.

1.4 Organization of Effort

1.4.1 The Structure of the Research Cores

The project is organized around three Core Research Groups: Applied Environmental Microbiology, Functional Genomics, and the Computational Core. The makeup of these groups is shown in Figure 1.10. Each group is responsible for one or more tasks, as outlined in Figure 1.9. The leader for each Core Group reports to the Director as outlined below, and is responsible for ensuring that project goals are achieved, work plans that support scientific



Applied Environmental Microbiology Core
LBNL, ORNL, U. Wash., UMC

Terry Hazen, LBNL, Group Leader
Alex Beliaeve, ORNL
Matthew Fields, ORNL
Hoi-Ying Holman, LBNL
Martin Keller, Diversa, Inc.
David Stahl, U. Washington
Dorthea Thompson, ORNL
Judy Wall, UMC
Jizhong Zhou, ORNL

Functional Genomics Core
LBNL, SNL, U. Wash., Diversa

Jay Keasling, LBNL/UCB, Group Leader
Alex Beliaeve, ORNL
Carolyn Bertozzi, UCB/HHMI/LBNL
Matthew Fields, ORNL
Martin Keller, Diversa, Inc.
Anup Singh, SNL
Dorthea Thompson, ORNL
Judy Wall, UMC
Jizhong Zhou, ORNL

Computational Core
LBNL, TIGR, U. Wash., UMC

Adam Arkin, LBNL/UCB, Group Leader
Inna Dubchak, LBNL
Frank Olken, LBNL

Conceptual Model Subgroup
Terry Hazen, LBNL
David Stahl, U. Wash.
Judy Wall, UMC
Jizhong Zhou, ORNL

Comparative Genomics Collaborator
Jonathan Eisen, TIGR

Figure 1.10. Composition of Core Research Groups

and technical objectives are implemented, data standards are developed and maintained, and efficient communication with the other core groups results in a seamless and powerful accomplishment of collaborative tasks. The three Core Research Group Leaders will report project status monthly to the Director and Steering Committee and make progress reports to the Technical Advisory Panel twice a year.

1.4.2 Project Management Structure

An Executive Committee, a Steering Committee, a Technical Advisory Panel, and a Scientific Advisory Committee will be instituted to ensure effective communications and management (Figure 1.11). The Executive Committee will provide executive leadership with wide-ranging authority to act in all facets of programs and personnel. Under the direction of the Project Director, the Steering Committee will manage scientific operations. The Technical Advisory Panel will offer independent guidance on the project's technical development and progress. A detailed chart of the functions, responsibilities, and authority of these governing bodies is provided in the Appendix.

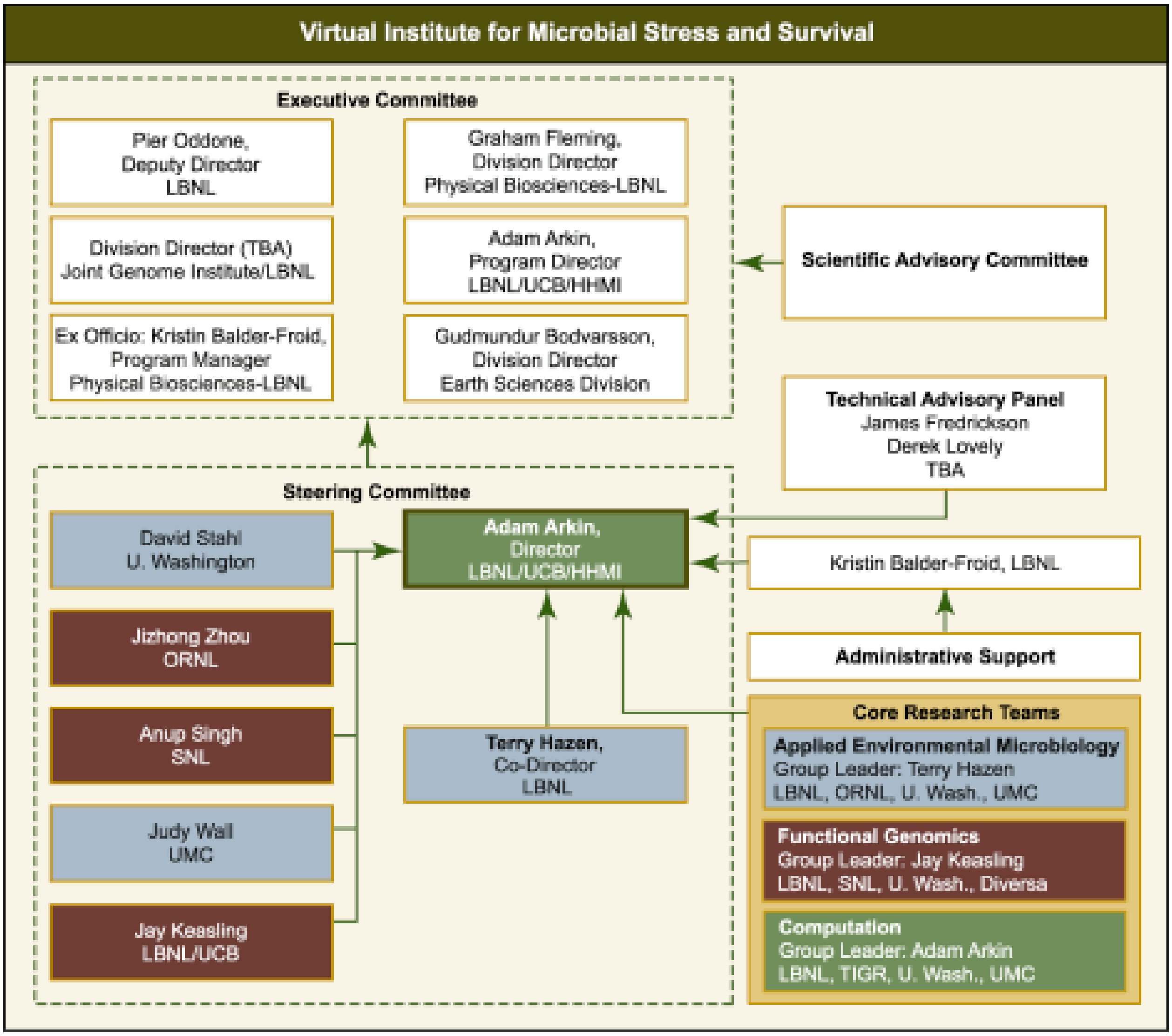

Virtual Institute for Microbial Stress and Survival
Executive Committee
Pier Oddone, Deputy Director LBNL
Graham Fleming, Division Director Physical Biosciences-LBNL
Division Director (TBA) Joint Genome Institute/LBNL
Adam Arkin, Program Director LBNL/UCB/HHMI
Ex Officio: Kristin Balder-Froid, Program Manager Physical Biosciences-LBNL
Gudmundur Bodvarsson, Division Director Earth Sciences Division
Scientific Advisory Committee
Technical Advisory Panel
James Fredrickson
Derek Lovely
TBA
Steering Committee
David Stahl U. Washington
Adam Arkin, Director LBNL/UCB/HHMI
Kristin Balder-Froid, LBNL
Jizhong Zhou ORNL
Administrative Support
Anup Singh SNL
Terry Hazen, Co-Director LBNL
Core Research Teams
Applied Environmental Microbiology
Group Leader: Terry Hazen
LBNL, ORNL, U. Wash., UMC
Judy Wall UMC
Functional Genomics
Group Leader: Jay Keasling
LBNL, SNL, U. Wash., Diversa
Jay Keasling LBNL/UCB
Computation
Group Leader: Adam Arkin
LBNL, TIGR, U. Wash., UMC

Figure 1.11. Composition of Core Research Groups.

The Executive Committee will be the project's highest-level internal governing body, meeting bimonthly to review the project's overall progress. This Committee approves annual work plans and major changes in scientific direction or operations. It also will have the authority to act in cases of conflict of interest or scientific nonperformance.

Comprised of representatives from each collaborating institution, the Steering Committee will be responsible for ensuring effective and efficient scientific operations. It will meet monthly (by video conference for non-LBNL personnel) to review resource allocations, budget performance, milestones, and timelines, and to assess progress on each task. The Steering Committee will also guide the development of project facilities, regularly reviewing work plans, general operations, and facility management. New directions and techniques that might arise from the facilities will be assessed, based on the advice of the Technical Advisory Panel. The Steering Committee will produce an annual "state of the project" report for the Executive Committee and DOE/BER management that will ultimately be posted on the project's Web site.

The Technical Advisory Panel will ensure the project's technical development is aligned with related DOE efforts working with the targeted organisms. The two individuals assigned to this Panel are James Fredrickson (who is involved in the Shewanella Federation) and Derek Lovely (who is working on a microbial cell grant on *Geobacter sulfurreducens*). Other members will be recruited for the Computational and Functional Genomic Core groups when work commences in those areas. The Technical Advisory Panel will be briefed monthly by the Steering Committee and will meet biannually with the committee (perhaps by video conference) to discuss the state of the work and possible changes in technical approach or biological focus, and to exchange data and information.

The Director and Co-Director are responsible for providing the overall scientific vision and technical goals and for ensuring that deliverables are accomplished. They will represent the project to the Department of Energy and be the main point of contact for the scientific collaborators. Effective communications and integrating project components will be their responsibility. They will oversee the activities of the Program Manager. They will also ensure that the collaborative functions of the Web site (in particular, database access and analytical tools) are maintained and available.

The LBNL Physical Biosciences Division Program Manager will develop plans for managing the growth of the project, assist the PIs and core research team leaders to develop effective communications, facilitate intellectual property issues, and oversee administrative support for the project. She will represent the project to DOE if the PIs are unavailable, and she will convene and provide support to the Executive Committee.

1.4.2.1. Reviews and Evaluations. A Scientific Advisory Committee will be convened by the Executive Committee to conduct an annual comprehensive review of the project's progress. Group membership will be comprised of DOE program managers and independent scientists with expertise in relevant scientific and technical areas. Every other year, the project will be included in the annual LBNL Director's Review of the Physical Biosciences Division, the host division of this project. The results of both of these reviews will be shared with DOE/BER. See Appendix A for a detailed breakdown of responsibilities and decision authority

1.4.3 Integration of the Program

The Core Research group leaders are responsible for ensuring the smooth operation of their section and cooperation with the other groups. They will be the main point of contact for collaborators to make their progress and desires known to the Director's Steering Committee and to the other group leaders. They will also ensure quality and compliance. For example, the Applied Environmental Microbiology leader is responsible for ensuring that sample preparation is standardized to the liking of the Functional Genomics and Computational Core Leaders. The Functional Genomics Leader will be responsible for ensuring quality control for the data production and timely introduction into the database. The Computational Leader will be responsible for ensuring that data entry, querying, and curation interfaces serve the needs of the other two groups, and that the models are usable by the biologists outside the modeling group. All three group leaders are responsible for seeing that data and tools are disseminated in a timely fashion. Finally, they are also responsible for overseeing the creation, hardening, and expansion of the user facilities that are a central aspect of this proposal. The physiological pipeline, the environmental simulation facility, the combinatorial chemistry facility, and the software tool and database technologies will be developed into reusable general-purpose facilities for the support of other DOE projects and the scientific community. Detailed descriptions of the facilities and growth plans are included at the end of each core research section.

1.4.4 Communication and Dissemination of Data and Tools

Communication and dissemination methods will follow customary practice and adhere to DOE guidelines. Results from the project will appear in peer-reviewed journals and in seminars given at major conferences and research centers. Technical reports and documentation on various special topics deemed inappropriate for individual publication will be made available on the project's Web site.

The Web site will be the public interface for the database and software tools developed by the group. The default policy is that all data, tools, and software developed during the course of this project will be made available, open-source, and free in a timely manner. Data will be released within 14 days, or at minimum, within a month. Tools and software will be available as they become usable, at minimum, by the end of the project period. Each piece of data and every tool developed will be evaluated for possible exceptions to this policy. If an exception is requested and the case seems reasonable, the issue will be brought to the Executive Committee for review, and then to DOE. If a limitation on release is then deemed appropriate, the data/tools may be made available only to the project researchers.

The Web site will also serve as an education and information dissemination tool for all of the findings and facilities of the project. This site will be modeled after the highly successful NABIR Program Web site that is operated by LBNL for DOE OBER (http://www.lbl.gov/NABIR/). The Web site, in addition to database and software tools, will have sections for each of the Core components, the latest publications, related links, news flashes, research plans, meetings/conferences, PI calendar, and private virtual lab notebooks so that investigators on the project can easily share their latest findings with the rest of the core teams, and descriptions of facilities and their contacts for other investigators as they come on line.

When the facilities and software become user-friendly, the directors will hold workshops and small conferences to train external scientists in the use of these tools. In addition, every two years, the Steering Committee will organize an international conference on stress response in environmental microorganisms. This will be a venue for top researchers in the area to share their new results with their professional community.

1.4.5 Resource Facilities: Implementation and Growth

Through the Virtual Institute for Microbial Stress and Survival (VIMSS) we also intend to develop five experimental core facilities that can be used by the VIMSS investigators and will develop into shared facilities for a variety of DOE funded projects and work for others. These core facilities included in VIMSS are: 1) the Environmental Molecular Microbiology Facility (ORNL, LBNL, U. Washington, and DIVERSA Inc.), 2) the Metabolite/Protein Profiling Facility (LBNL, U.C. Berkeley), 3) the Combinatorial Chemistry for Functional Genomics Facility (LBNL, U.C. Berkeley), 4) Computational Biology Facility (LBNL), and 5) Microbial Genomics Laboratory (ORNL). The implementation and growth of these facilities are described in the core research sections.

1.4.6 Milestones Summary

Lawrence Berkeley National Laboratory, Oak Ridge National Laboratory, Diversa, Inc.

University of California, Berkeley, University of Missouri at Columbia, University of Washington

Year 1 (July - September, 2002): Begin survey and mapping of ORNL sites for stressors and target organisms. Begin development of environmental chambers. Start expression profiling experiments with *S. oneidensis*. Complete whole-genome microarray construction for *S. oneidensis*. Design whole-genome microarrays for

D.vulgaris and *G. metallireducens*. Compilation of homologous stress response pathways. Creation of version 1 of the comparative genomics pipeline (CGP). Application to target organisms.

Year 2 (October, 2002 – September 2003): Complete survey and mapping of ORNL site and begin survey and mapping at LLNL site. Begin cell production for Functional Genomics Core. Provide field and lab data to Computational Core. Complete expression profiling experiments with *S. oneidensis.* Generation of deletion mutants and chemical inhibitors of *S. oneidensis*. Whole-genome microarray construction for *D.vulgaris* and *G. metallireducens.* Flat-file databases implemented. Relational schema designed. Graph database implementation designed. CGP and VISTA integrated. *S. oneidensis* pathways predicted. Modeling begins.

Year 3 (October, 2003 – September 2004): Complete development of environmental simulators and chemostats. Complete stressor studies in chemostats and soil columns for expression analysis, PLFA, SFTIR, etc. Complete expression profiling experiments with regulatory mutants of *S. oneidensis. Complete e*xpression profiling experiments with *D.vulgaris* and *G. metallireducens.* Transcript, protein, and metabolite profiling of wild-type, deletion mutants, and chemically inhibited *S. oneidensis.* Generation of deletion mutants and chemical inhibitors of *G. metallireducens.* Relational database implemented. *S. oneidensis* model predictions tested against new data. Reverse engineering of *G. metallireducens* and *D. vulgaris* accomplished and models begun. CGP version 2 released. Regulatory strategies clustered.

Year 4 (October, 2004 – September 2005): Complete stressor studies in chemostats and soil columns for expression analysis, PLFA, SFTIR, etc. Determine incidence of closely related relatives to target bacteria in chemostats, soil columns and field samples. Continue cell production for Functional Genomics Core. Complete expression profiling experiments with *D.vulgaris* and *G. metallireducens.* Transcript, protein, and metabolite profiling of wild-type, deletion mutants, and chemically inhibited *G. metallireducens.* Determine protein-protein interactions in S. *oneidensis*. Graph database implemented and final interfaces released. Comparative modeling of the three organisms begin. Release of CGP tuned for stress response in these three organisms. First validated models delivered to AEMC and FGC for conceptual modeling.

Year 5 (October, 2005 – September 2006): Continue cell production for Functional Genomics Core. Begin testing of conceptual models in both the laboratory and field using new probes and assays. Complete expression profiling experiments with regulatory mutants of *D.vulgaris* and *G. metallireducens.* Generation of deletion mutants and chemical inhibitors of *D.vulgaris*. Determine protein-protein interactions in *G. metallireducens.* Database schema and software released. Comparative models of targets released. Prediction of significant non-target organisms in need of further experimental study (by regulatory clustering and modeling). Combined population models created.

Year 6 (October, 2006 – September 2007): Complete testing of conceptual models in both the laboratory and field using new probes and assays. Complete survey and mapping of other contaminated DOE sites for verification of general applicability of conceptual models using new probes and assays. Transcript, protein, and metabolite profiling of wild-type, deletion mutants, and chemically inhibited *D.vulgaris*. Determine protein-protein interactions in *D. vulgaris*. Single and combined models further validated against data. Conceptual models released and tested. Final software, database and documentation release.

2 APPLIED ENVIRONMENTAL MICROBIOLOGY CORE

Terry C. Hazen (LBNL, Core Team Leader)
Alex Beliave (ORNL)
Matthew Fields (ORNL)
Hoi-Ying Holman (LBNL)
David Stahl (University of Washington)
Judy Wall (University of Missouri)
Jizhong Zhou (ORNL)
Martin Keller (Diversa Inc.)

2.1 Goals and Specific Aims

The main goal of the proposed research is to develop criteria for monitoring the integrity (health) and altering the trajectory of an environmental biological system (process control). To achieve this requires a more complete understanding of how the biological "units" comprising the system are organized, regulated, and linked in time and space (genes, genomes, cells, populations, communities and, ultimately, ecosystems). Key to these objectives is a more complete understanding of the diversity and environmental context of stress response systems. To accomplish these goals, the overall project has three distinct science cores that are dependent on each other for products, outcomes, feedback, and verification of results and models (Figure 1.9). The Applied Environmental Microbiology (AEM) Core is the source of environmental data and samples that determine the stressors that will be studied, provides the environments for growing the organisms to be tested, simulates stressed environments, and verifies the conceptual models to determine how these stress regulatory pathways control the biogeochemistry of contaminated sites (Figure 2.1). The specific goals of the AEM Core are to

1. Survey and map DOE sites contaminated by metals and radionuclides using chemical and molecular/ microbiological parameters to determine major microbial populations and potential stressors for *Desulfovibrio vulgaris*, *Geobacter metallireducens*, and *Shewanella oneidensis* MR1.
2. Determine the rank priority of these stressors in terms of their ability to affect metal/radionuclide bioreduction by either direct or indirect processes, and to establish the normal active range of the stressor in metal/radionuclide contaminated environments.
3. Determine the incidence and activity of the three target bacteria, and closely related relatives, in the test metal/radionuclide contaminated environments and collect isolates for analysis by the Functional Genomics Core for comparison using 16S RNA profiling.
4. Create defensible environmental simulators that can replicate key features of field site chemical and biological structure to mimic stress conditions for single populations and later for microbial communities (chemostats to soil columns from 10 µm to 1 m size systems).
5. Provide large quantities of cells in various stress states for the Functional Genomics Core's physiological monitoring facility, molecular interaction studies, and combinatorial chemistry group.
6. Provide environmental simulators for testing stressor effects on mutants, large insert clones, expression analysis, etc., for elucidating critical parts of the stress regulation pathway.
7. Develop testable conceptual models of stress regulatory pathways based on results of the Computational Core that could predict natural attenuation and suggest biostimulatory strategies for immobilization of metals and radionuclides at DOE contaminated sites.
8. Test conceptual models of stress regulatory pathways and effects on contaminate site biogeochemistry using competent soil columns with different levels of complexity over the active range of the stressors
9. Validate conceptual models using field tests at contaminated sites that utilize specific functional gene arrays developed from the stress regulatory pathways.
10. Alter field conditions or test along gradients to verify stress regulatory model efficiency for predicting natural attenuation or suggesting biostimulatory strategies for immobilization of metals and radionuclides.

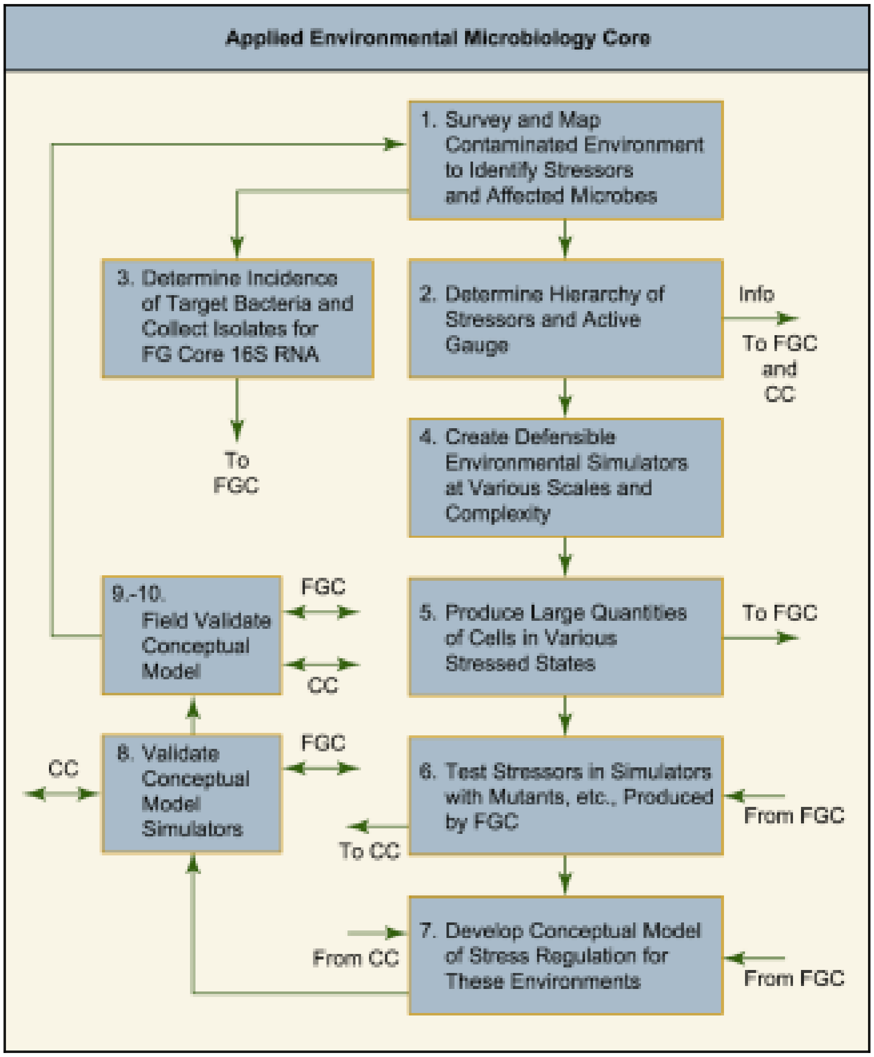

Applied Environmental Microbiology Core
1. Survey and Map Contaminated Environment to Identify Stressors and Affected Microbes
2. Determine Hierarchy of Stressors and Active Gauge
Info
To FGC and CC
3. Determine Incidence of Target Bacteria and Collect Isolates for FG Core 16S RNA
To FGC
4. Create Defensible Environmental Simulators at Various Scales and Complexity
5. Produce Large Quantities of Cells in Various Stressed States
To FGC
9.-10. Field Validate Conceptual Model
FGC
CC
8. Validate Conceptual Model Simulators
CC
FGC
6. Test Stressors in Simulators with Mutants, etc., Produced by FGC
To CC
From FGC
7. Develop Conceptual Model of Stress Regulation for These Environments
From CC
From FGC

Figure 2.1. Specific goals for the AEM Core and where they input from and output to the FGC and CC Cores.

2.2 Background and Significance

2.2.1 Environmental Considerations: An Ecological Framework for Defining Stress

Extensive application of direct molecular measures of microbial community structure in different environments/habitats (primarily based on comparative 16S rRNA sequencing) has revealed some general themes that are directly relevant to the goals of the proposed research. First, high population diversity is common, and the greater part of environmental diversity has yet to be recovered in pure culture. Thus, research leading to an applied endpoint in open environmental systems must consider this dominant component of the system.

Second, there is a natural organization to microbial diversity within a given habitat. Generally, populations (as identified by the 16S rRNA sequence type) are affiliated with well-resolved phylogenetic groups (clades or "phylotypes") (5, 29). For example, inspections of sediment systems in which metal or sulfate reduction is a major process have revealed a high diversity of related populations (e.g., members of Geobacteraceae and Desulfovibrionaceae) (12). Thus, at higher taxonomic ranks there is a less complex natural order—high species diversity is captured by a limited number of major phylotypes. This suggests that the natural assemblages share a similar ecological function (e.g., metal reduction or sulfate reduction).

However, there is still little understanding of the ecological factors contributing to, or sustaining, high diversity within a phylotype (e.g., within the families Geobacteraceae or Desulfovibrionaceae). We suggest that high species diversity among populations comprising a single phylotype contribute to functional stability within an environmental system. Since environments are not static but are constantly changing with respect to key physical-chemical variables (e.g., substrate concentration, temperature, pH, salinity, osmolarity, light, redox potential, etc.), it follows that not all populations in a system are simultaneously experiencing optimal growth. Those populations growing under suboptimal conditions, or entering resistant (e.g., spores) or moribund states, are experiencing stress. Thus, we suggest that it is not sufficient to monitor stress. It is essential to understand the environmental context of stress response for different populations. We hypothesize that within any well-adapted and dynamic system, some fraction of the community will be experiencing stress. The key question is, What levels of stress response signal a system that is in distress or that is not optimally adapted?

2.2.2 Issues of Scale for Stress and Biotransformation

The success that anaerobic microsite models have had in explaining denitrification within macroscopically aerobic soils suggests that this perspective can be extended to redox biotransformations of contaminants and stress. Microbial respiration and diffusion-limited oxygen transport within soil aggregates permit development of local redox gradients. Even when drained macropores are nearly in equilibrium with atmospheric oxygen, nitrate reduction can proceed within interior regions of soil aggregates when the rate of diffusive oxygen supply cannot keep up with microbial community respiration rates (8, 28, 48, 60). Besides nitrogen, the coupling of microbial respiration, diffusion-limited oxygen supply, and intra-aggregate redox zonation has relevance to the fate of other redox-sensitive elements, including chlorinated organics (10), and trace element contaminants, such as Se (50, 51) and Cr (52). Given sufficiently large diffusion distances and respiration rates, the full spectrum of redox conditions relevant to contaminant transformations in some field sites can be contained within individual soil aggregates. Recent studies on soil aggregates have revealed additional insights into mesoscale heterogeneity, such as higher microbial activity along outer regions of larger aggregates (11) and greater enrichment with respect to recently deposited carbon on surfaces of larger aggregates (2, 10). Laboratory reactor and soil column studies clearly need to include different size systems in order to capture a more realistic appraisal of metal and radionuclide contaminated sites.

2.2.3 Why Desulfovibrio?

The sulfate-reducing bacteria (SRB) are classified by only two characteristics: their oxygen sensitivity and their ability to use sulfate as a terminal electron acceptor. This sweeping classification includes many types of bacteria, gram negative and positive, mesophilic and thermophilic, marine and freshwater, and eubacteria and archaea. However, members of the genus *Desulfovibrio* are the most readily cultured and are the only SRB that have been subjected to molecular biological analysis. *Desulfovibrio vulgaris* subsp. *vulgaris* strain Hildenborough has been the SRB of choice for many biochemical studies as well as the first SRB to be examined by molecular tools (40, 56). It now claims the distinction of being the first SRB to be fully sequenced [data available from The Institute for Genome Research (TIGR)]. Another SRB, *Desulfovibrio desulfuricans* strain G20, is currently being sequenced. Because its physiology is quite similar to *D. vulgaris*, comparisons of responses and genes will help identify important global stress responses.

All the historical interfaces of the SRB with human activities result from the energy-generating processes of these anaerobes. While many proteins that are candidates for bioenergetic pathways have been studied (39), the connections and complete pathways remain to be resolved. The extent of the metabolic versatility has recently been expanded. The anecdotal observations of these microbes in aerobic environmental niches are now being supported by data on the multiple systems for the reduction of oxygen seen in the genome sequence and are being experimentally demonstrated (9, 16, 33). Thus, the functional habitats of these “strictly anaerobic” bacteria are likely to be much wider than previously considered. More recent studies (14, 32) have documented the ability of a number of *Desulfovibrio* strains to reduce toxic metals, such as uranium and chromium, a process that results in the production of less-water-soluble species. A physiological role for the reduction of toxic metals by the bacteria awaits elucidation. Stress response pathways are sure to play an important role in the niche definition of *Desulfovibrio* and its effect on the biogeochemistry of many contaminated environments.

2.3 Preliminary Studies

2.3.1 Reductase Genes and Geochemical Parameters

Microorganisms are often exposed to multiple stress conditions in situ, and the pleotrophic effects on community structure and functional gene pools are most likely mediated via the stress-response systems of individual microorganisms. Our recent results (ORNL) suggest that definitive relationships between geochemical parameters and microbial communities are difficult to elucidate (60). When the distribution of cloned nitrite reductase genes, nirK and nirS, were compared to geochemical measurements (levels of nitrate, uranium, pH, heavy metals, and organic carbon) at six groundwater sites, the cloned genes appeared to cluster in relation to nitrate and nickel levels. Component analysis also indicated that pH was a strong negative loading factor for contribution to overall variability. However, the types of genes (sequence diversity) could not be predicted with high confidence when a linear or logistic model was used. An artificial neural network improved the prediction ability of sequence distribution, and this result indicated that the relationships between microorganisms with a nirK or nirS gene and site geochemistry were nonlinear. In addition, diversity indices for the gene pools at acidic contaminated sites were sometimes higher than the background site, and the same predominant sequence was observed at the acidic contaminated sites as well as the background site (Table 2.1). These results indicate that the effects of multiple stresses on microbial communities and functional capacities are complicated and not well understood. We propose to study the effects of relevant stress conditions in three environmentally important, radionuclide-reducing bacteria: *Desulfovibrio vulgaris, Geobacter metallireducens,* and *Shewanella oneidensis*.

Table 2.1. Physical characteristics and diversity estimates for the *nir*K and *nir*S gene clones from six FRC groundwater samples. FW-300 represents the background area, FW-003 has circum neutral pH but high nitrate (1,000 ppm), TPB-16 has circum neutral pH, low nitrate (30 ppm), and the acidic sites, FW-005, FW-010, and FW-015, have low pH (3.5–3.9), high nitrate (100–40,000 ppm), high nickel, and high uranium.

	%Coverage	H'[a]	1/D[b]	Evenness[c]	Richness[d]	# of clones[e]
***nir*K**						
FW-300	89	2.33	3.30	0.58	22±5	143
FW-003	95	0.70	1.20	0.20	37±19	229
FW-005	94	0.78	1.23	0.22	18±5	185
FW-010	91	2.13	2.90	0.60	18±6	134
FW-015	85	3.10	5.11	0.71	71±33	140
TPB-16	78	2.93	4.07	0.67	35±9	127
nirS						
FW-300	80	2.92	3.10	0.55	84±18	173
FW-003	78	3.86	6.51	0.75	39±3	210
FW-005	77	3.60	4.01	0.61	243±58	253
FW-010	74	2.15	2.59	0.52	38±14	174
FW-015	80	1.33	1.47	0.32	27±7	175
TPB-16	81	1.91	2.07	0.45	28±7	177

[a] Sannon-Wiener index, higher number represents more diversity

[b] Reciprocal of Simpson's index, higher number represents more diversity

[c] As Evenness approaches 1, the population is more evenly distributed

[d] Statistical prediction of the number of different species or genes

[e] Number of clones in each sublibrary analyzed by RFLP

2.3.2 Microenvironments

Our recent studies have shown that the fate of chromium contamination within soil aggregates can be strongly diffusion-limited, resulting in reduction to Cr(III) within short distances (52). In large diffusion-limited domains, the Cr-contamination can be restricted to outer regions in contact with preferential flow paths, leaving the deeper core region unaffected. Such aggregates contain microbial populations that have and have not been exposed to Cr(VI), which reside within outer and core regions, respectively (Figure 2.2). These results show the importance of intra-aggregate spatial relations for redox-sensitive contaminants as well as for the microbial communities responsible for producing redox gradients and reductants. By extension, we anticipate that similar stratification of redox potentials, metal contaminants, and microbial communities will occur within larger sediment blocks deeper in the subsurface. In soils and sediments comprised of aggregates or blocks that support internal redox gradients, bulk characterization of chemical and microbiological characteristics does not allow mechanistic understanding of biogeochemical processes. Communities, not total biomass, control net process rates, driving biogeochemical cycles and the transformation of pollutants. Thus, descriptions of the temporal and spatial dimensions of microbial community structure and the complex gene expression patterns that underlie trophic interactions and stress response are fundamental to a more complete understanding of environmental processes. Stress regulatory responses of the bacteria to variable environments will strongly control the rate, extent, and location of metal reduction in the subsurface.

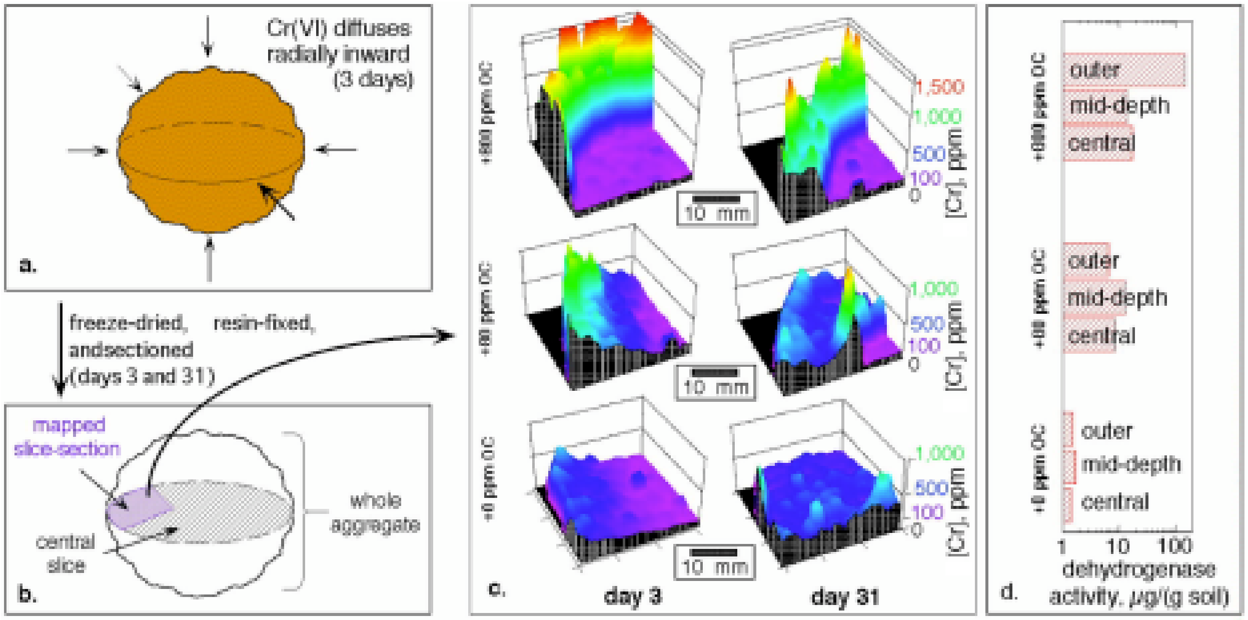


Figure 2.2. Synchrotron x-ray microprobe maps of total Cr in sections from 6 soil aggregates that were exposed to Cr(VI). The more reducing conditions established in the higher organic carbon aggregates result in more rapid Cr(VI) reduction at shorter distances, higher diffusive uptake of Cr, sharply terminated contaminant fronts, and higher microbial activity as indicated by dehydrogenase activity (52).

2.3.3 Synchrotron IR Probing of Living Bacteria Reducing Chromium

LBNL's newly developed synchrotron radiation-based (SR) Fourier-transform infrared (FTIR) spectromicroscopy beamline allows the study of the evolution of biogeochemical phenomena on heterogeneous surfaces of geological materials. For example, one can study the same bacterial colony and the same mineral nodules over time, which will eliminate experimental uncertainties induced by intercolony or mineral differences. The principle of the technique is straightforward. Different materials (chemicals, microbes, minerals, transformation products) absorb different wavelengths of infrared light; therefore, light transmitted or reflected by a sample yields a unique

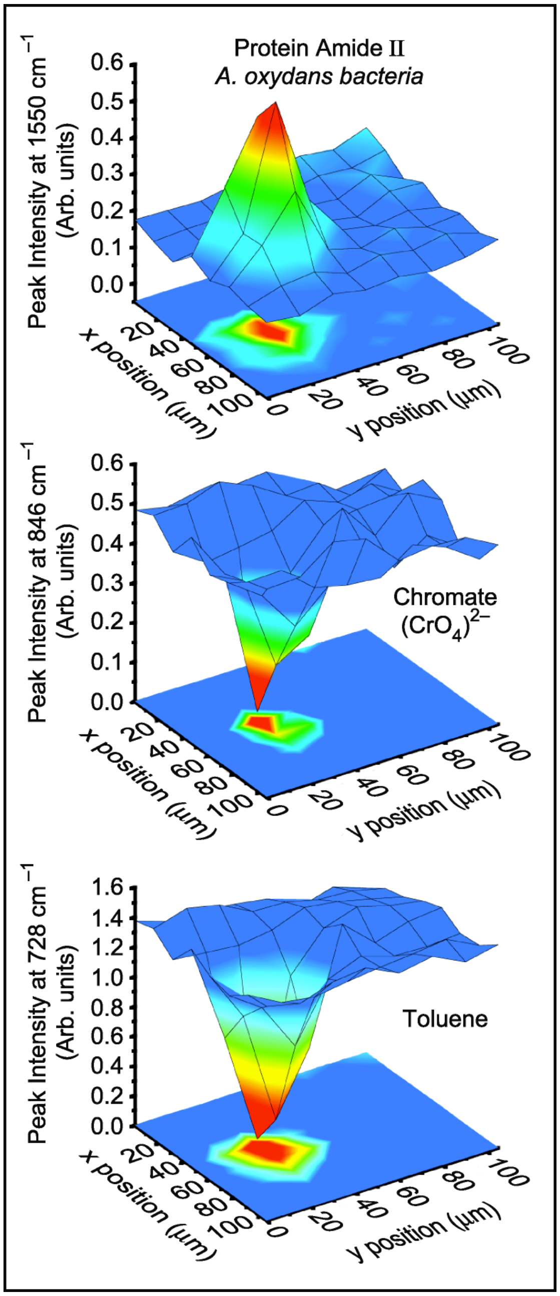


Figure 2.3. SR-FTIR mapping showing *Arthrobacter oxydans* bioremediating mixtures of Cr^{6+} and toluene on a mineral surface (23).

spectrum that can distinguish among different materials and among those undergoing different chemical and physical changes. Key to success is the quality of the infrared light source. Synchrotron light from the infrared beamline at the Advanced Light Source is used because it can be focused to a spot less than 10 m in diameter with hundreds of times the signal-to-noise ratio of conventional infrared sources. Although geological materials are inherently heterogeneous and have low reflectivity surfaces, the high brightness of the synchrotron beam enables one to measure, identify, and locate key compounds undergoing changes, as well as microbes involved in the changes. The technique has been successfully used to identify a new pathway of chromate reduction on surfaces of geological materials (23, 27, 59) and to provide direct insight into the relationship between localization of chromate reduction, toluene transformation, and *Arthrobacter oxydans* bacteria distribution on surfaces of mixed-valence iron oxides (Figure 2.3).

2.4 Research Design and Methods

Our general design philosophy will be to develop laboratory reactor systems that replicate key features of the microbial population structure and processes at contaminated sites as a prelude to conducting experiments to identify the spectrum of natural stressors. We emphasize that it is essential to identify stress in a natural context prior to evaluating stress response in less complex reactor communities and simplified systems containing our key target organism (*Desulfovibrio vulgaris*).

Our general research structure is composed of five main elements: (1) chemical and molecular/microbiological mapping of field sites to identify major microbial populations and potential system stressors; (2) development of reactor systems that replicate key features of field site chemical and biological structure; (3) testing of different stressors using appropriate controlled reactor systems in combination with a general measure of physiological stress (rRNA synthesis); (4) development of simplified reactor systems containing the target organism, mutant variants, and native populations that are closely related to the target organism; and (5) field validation of conceptual models using chemical and molecular/microbiological probes and assays based on the stress regulatory pathways developed during the overall project. Our past studies of bioreactor systems have shown that it is possible to develop complex communities containing populations closely related to *Desulfovibrio vulgaris* and *Geobacter* spp. We will build upon these earlier reactor system studies in the development of appropriate systems for the proposed research.

2.4.1 Integration with Other Research Cores

We will work closely with the other two GTL Cores on all aspects of these studies, sharing information of specific stressors and stress response characteristics of different populations comprising reactor and defined culture systems. This integration is key to the general project objective of establishing how pathways are regulated differently among different organisms because of niche constraints, evolutionary divergence of pathways, or lateral inheritance. We will establish reactors of sufficient scale and replication to provide ample material for coordinated physiological and molecular studies. For example, one objective of developing reactor systems that sustain native populations of varying relationship to our target organisms is to retrieve homologous pathways, asking if they serve the same roles in stress response for closely related, as well as for more distantly related, populations. This perspective is critical to the overall project objective of using information derived from laboratory studies to evaluate the status of natural environmental systems. Natural systems are comprised of tremendous phylogenetic diversity and have large representations of closely related populations (12). In addition, there is good comparative evidence that stress response systems that appear to be orthologous by comparative sequencing may serve different roles. For example, the *Caulobacter* sigma factor orthologous to the nitrogen-regulating sigma subunit RpoN of *E. coli* is required for stalk biogenesis and for normal cell division (5).

We will also work closely with the Functional Genomics Core and the Computational Core to link stress response pathways (identified in DNA recovered from reactor systems by comparative genomics/phylogenetics) with specific populations resident in reactor systems and field sites. The linking of genes with responsive populations is also a critical element of our general research strategy. This linking will be facilitated by a combination of new methods, using a radiomicroarray to identify stress responsive populations and high-speed cell sorting to associate stress responsive populations with specific genes.

2.4.2 Chemical and Molecular Microbiological Mapping of Field Sites

2.4.2.1. Field Sites for Sampling. Initially two DOE sites will be used for study, the Oak Ridge NABIR Field Research Center and the Pit 7 Complex at Site 300 of LLNL. Both sites are well characterized and have uranium as a principal contaminant of concern. The Oak Ridge site has high concentrations of uranium, low pH, and high nitrate. The LLNL site has much lower concentrations of uranium, neutral pH, and low nitrate. Samples can be obtained from Oak Ridge via the NABIR Field Research Center and from the LLNL site restoration manager. The three Core Groups also have access to several Office of Environmental Management funded projects at metal and radionuclide contaminated sites at Savannah River, Hanford, and INEEL, which could be sources for samples for later studies.

NABIR Field Research Center at Oak Ridge. The NABIR Field Research Center at ORNL has both a contaminated and a control area. Actinide contamination at this site resulted from disposal of uranium processing waste to the nearby S-3 Waste Disposal Ponds and to the Boneyard/Burnyard (BY/BY). The S-3 Waste Disposal Ponds consisted of four unlined surface impoundments that were constructed in 1951 (59). They received liquid nitric acid and uranium-bearing wastes via a pipeline at a rate of approximately 10 million liters per year until 1983. The Ponds were unlined and approximately 122 m _122 m wide _5.2 m deep. Infiltration was the primary release mechanism to soils and groundwater. The S-3 Ponds were neutralized and biodenitrified in 1984, and subsequently closed and capped in 1988. The BY/BY is not located within the proposed contaminated field research area but is one of the primary sources of uranium contamination. The BY/BY includes (1) the Boneyard, which consisted of unlined shallow trenches used to dispose of construction debris and to burn magnesium chips and wood; (2) the Burnyard, which was used from 1943 to 1968, and received wastes, metal shavings, solvents, oils, and laboratory chemicals that were burned in two unlined trenches; and (3) the Hazardous Chemical Disposal Area (HCDA), which was built over the Burnyard and handled compressed gas cylinders and reactive chemicals. The residues from the cylinders and reactive chemicals were placed in a small, unlined pit. Although the HCDA has been capped, the rest of the BY/BY remains uncapped. This has resulted in U contamination along with small amounts of Np and Pu in both the vadose zone and saturated zone in the area.

The original source material had significant amounts of organics and high concentrations of nitrate, which created anaerobic and reducing conditions. Since there are no new sources of nitrates and organics, these contaminants have been depleted over time, and thus the subsurface is gradually reverting to oxic conditions. Background concentrations of dissolved oxygen in the residuum at the reactive barrier sites are generally 2 ppm, but vary between 1 and 4 ppm. Eh varies between 100 and 300 mV, and pH varies between 5.0 and 6.5. The terminal electron acceptor process in the shallow residuum is likely driven by oxygen. Deep groundwater in the Nolichucky Shale and Maynardville Limestone can be anaerobic and have a negative Eh (< –250 mV). Under these groundwater conditions, nitrate is likely to be the most important to the electron acceptor process. Reactive barriers (i.e., subsurface iron walls) at the S-3 Ponds site were installed using guar gum. After the guar was broken down by injecting an enzyme, microbial activity in the gravel-filled trenches associated with the barriers increased dramatically, particularly with sulfur- and iron-reducing microbes. Pre-guar gum concentrations of nitrate (>1,000 ppm) and uranium (>2 ppm) in these trenches were reduced dramatically to the low ppb level after the guar gum was injected, suggesting the potential for removal of these contaminants through microbial activity. Although groundwater U concentrations have been reduced, most likely by microbial reduction, the conditions in similar subsurface areas at this site are oxic. Therefore, it is reasonable to expect that oxia will return to this system over time.

The ORNL site represents a typical scenario for uranium contamination at DOE sites, and as such, is relevant to our proposed studies. Indeed, anaerobic conditions existed in the past (and in deeper areas currently), thus creating the reducing conditions that would transform U. These environments will provide us an array of environmental uranium-contamination conditions from which we can obtain samples to test our hypotheses.

Pit 7 Complex at Site 300 of LLNL. The Pit 7 Complex at Site 300 of Lawrence Livermore National Laboratory is contaminated with uranium, tritium, and trace amounts of other chemicals. The principal contaminants of concern are tritium and uranium. Tritium can be remediated by hydraulic control and natural attenuation since it has a fairly short half-life. Uranium presents a much more serious problem since its half-life does not allow natural attenuation to be considered as a strategy. LLNL currently has no solution to this problem. The uranium contamination at the Pit 7 Complex came from depleted uranium components that were exploded on the firing range between 1958 and 1989. Firing table gravel and debris from the Building 850 firing table were disposed of in the Pits 3 and 5 landfills of the Pit 7 Complex. Leaching from these unlined landfills has resulted in the release of VOCs, uranium, and tritium to the subsurface. The maximum concentration of uranium (187 pCi/L) in groundwater near the pits was detected in samples collected in 1998. Data indicate that the extent of uranium-238 in groundwater is limited primarily to alluvium immediately adjacent to the pits; however, it has also been detected in the underlying sandstone bedrock (Figure 2.4). The pits were constructed by excavating topsoil and alluvial material to an average depth of 18 to 20 ft. The Pit 3 Landfill contains approximately 26,200 yd^3 of material contained in a 6,200 yd^2 area. The Pit 5 Landfill contains approximately 29,920 yd^3 of waste material in a 9,100 yd^2 area. The pits are underlain by sandstone bedrock with siltstone/claystone interbeds. Depth to ground-water is approximately 15–25 ft below the valley bottom where pits are located. During heavy rainfall/groundwater periods, water levels rise into the pits or enter the pits laterally. The uranium currently contaminates 1,500,000 ft^3 of ground water, covering an area of more than 5,000,000 ft^2.

The LLNL site also represents a fairly typical scenario for depleted uranium contamination at DOE sites and, as such, is relevant to our proposed studies. Indeed, episodic infiltration of the pits leaches uranium into the subsurface and is further complicated by the aggregate nature of the alluvium and the siltstone/claystone interbeds of the underlying sandstone bedrock. These environments will provide us an array of nearby environmental uranium-contamination conditions from which we can obtain samples to test our hypotheses. Our previous studies on aggregates from the Altamont Hills are directly adjacent to this site. This site also provides a good contrast to the high-nitrate sites, such as the FRC, and will allow us to determine if the microbial community adapted under these low-nitrate conditions will respond to stresses differently.

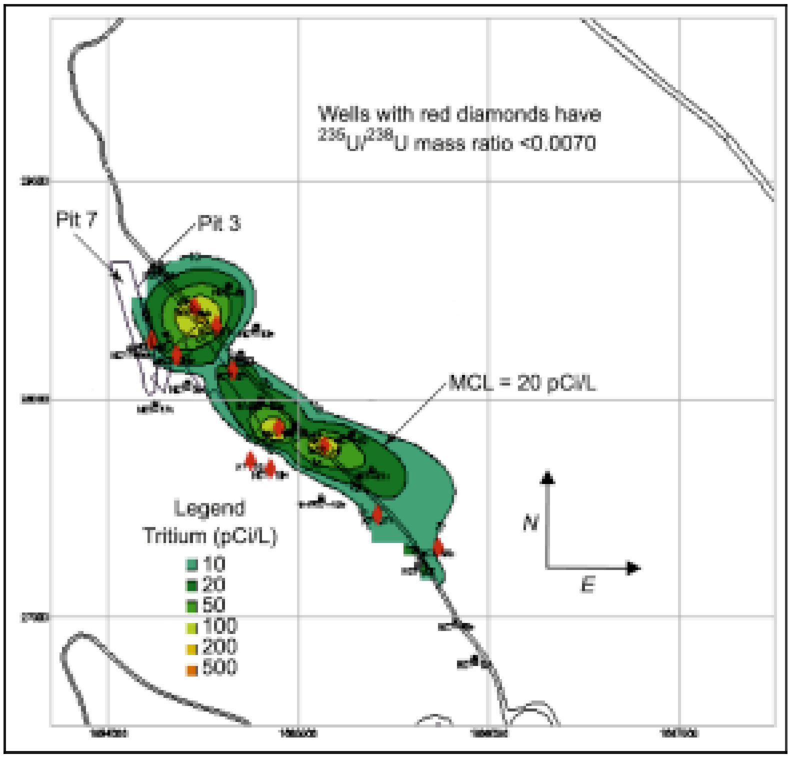

Wells with red diamonds have
$^{235}U/^{238}U$ mass ratio <0.0070
Pit 7
Pit 3
MCL = 20 pCi/L
Legend
Tritium (pCi/L)
10
20
50
100
200
500
N
E

Figure 2.4. Map showing tritium and uranium distribution at Site 30 Pit 7, Lawrence Livermore National Laboratory.

2.4.2.2. General Sediment and Soil Sampling. Sediment and soil samples will be recovered from these sites along transects of increasing levels of contaminants/stressors (pH, nitrate, oxygen, and uranium concentrations). Replicate samples will be taken at intervals of sufficient density to capture major trends in chemical structure. Sampling strategy may be modified in consideration of results obtained from initial mapping of microbial population structure. For example, if we observe major shifts in community structure within replicate samples at one point along a sampling transect, we would revise our sampling protocol to better associate site chemistry with community structure. We anticipate that 5–20 grams of site material will be sufficient for initial mapping of a population structure based on the recovery of DNA for a molecular assessment based on 16S rRNA sequence diversity. The initial survey will be based on a terminal restriction fragment length polymorphism (T-RFLP) of amplified 16S rRNA sequences (31). Investigations already being conducted at the FRC by several NABIR investigators and the ORNL group associated with this project will enable us to do these studies with a minimum of preliminary characterization. (For more information on the NABIR FRC, see http://www.esd.ornl.gov/nabirfrc/.)

2.4.2.3. Chemical Analyses of Field Sites. Specific methods will depend upon the nature of the sediments collected, varying between sandy and clay/humic. Interstitial pore water samples will be preserved by acidification (cations and metal analyses), freezing (nutrients, except SiO_{2d}), or refrigeration (anions). The concentrations of unstable species (e.g., sulfides) will be determined directly in the field using colorimetric methods. Pore waters and sediments will be analyzed for major ions using a collection of methods that have been used previously by members of the AEM Core. The LLNL Site 300 is within two hours' driving time of our labs at LBNL, where we have an EPA-certified environmental measurements laboratory that can analyze all anions, cations, volatiles, metals, radionuclides, trace nutrients, and even stable isotopes. The ORNL NABIR Field Research Center has similar facilities and has already sampled this site extensively for an extremely wide array of chemical species (http://www.esd.ornl.gov/nabirfrc/). Both the ORNL and LBNL laboratories have the capability to analyze any type of sample that we might need.

2.4.2.4. T-RFLP Population Mapping. T-RFLP offers a reproducible way to rapidly describe population structures based upon sequence length polymorphisms in rapidly evolving regions of small subunit rRNAs (30). In a typical experiment, one or both primers in a PCR amplification are derivatized with a fluorescent ligand at the 5′

terminus. Only the terminal fragments are labeled in a restriction digest of the PCR products, and these can be resolved on a DNA sequencing gel. This method has been used recently to characterize the microbial population structure in sediments (4, 54). We will use T-RFLP mapping to refine our understanding of the variability of microbial populations at relatively high spatial resolution along gradients of varying contaminant and site chemistry. Identified changes in peak patterns will indicate shifts in microbial populations. Initial T-RFLP studies will use primers designed for general amplification of members of the bacterial domain (4, 49). We will also conduct a more limited survey using archaeal primers, but since this group is generally much less abundant than the bacterial domain, we will expand these studies only if notable population trends are observed in relationship to site chemistry.

2.4.3 Reactor System Development

We will evaluate three different reactor formats for the proposed studies of stress response and genetics: (1) fixed-bed reactors (Figure 2.5a), (2) fluidized bed reactors (Figure 2.6), and (3) soil columns (syringes to 2 m □ 9 cm) (Figure 2.7). All three reactor configurations provide for the development of complex attached communities representative of the more common state of growth of microorganisms in sediment and soil habitats. Our past analyses of a fixed-bed reactor inoculated from groundwater demonstrated that large populations of bacteria are closely related to our key target organism (*Desulfovibrio vulgaris* subsp. *vulgaris* Hildenborough) could be selected for and maintained under a range of reactor operating conditions (Figure 2.9) (44). Most notably, *Desulfovibrio* sp. were abundant under operating conditions of both low and high sulfate, comprising about 20% and 30% of the community under those two operating conditions (Figure 2.5b)(44). This observation further supports our changing appreciation of the metabolic versatility of sulfate-reducing bacteria.

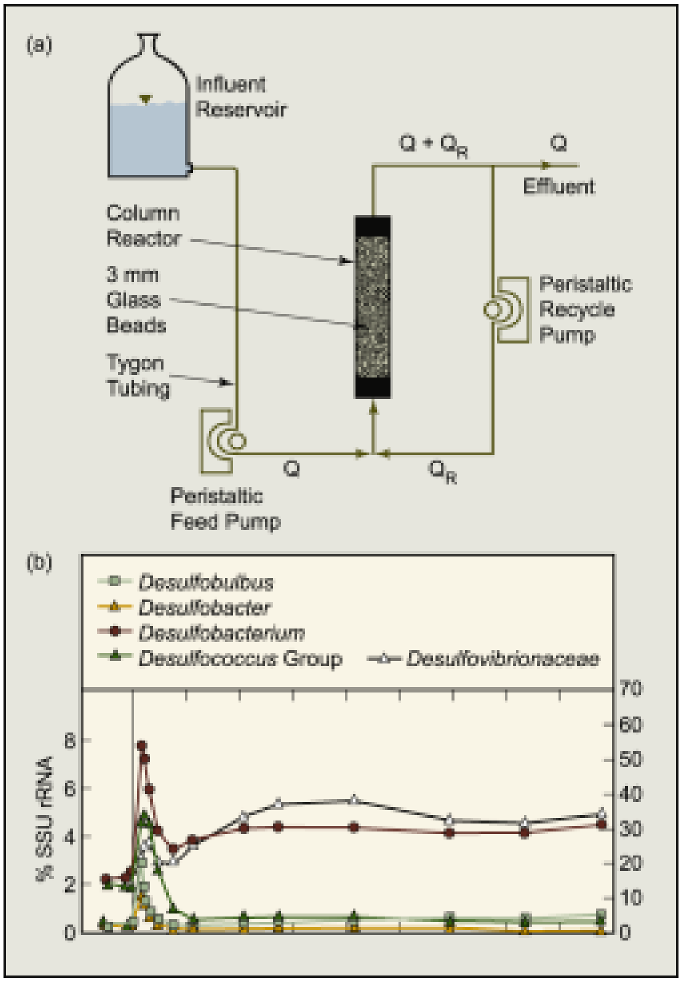


Figure 2.5. (a) Top. Fixed-bed complete-mix bioreactor (47). (b) Bottom. Microbial structure of fixed-bed bioreactor system operated at low (before day 0, indicated by vertical line) and high sulfate addition to the reactor feed. Each hatch mark represent 10 days on the time axis. Major populations of sulfate-reducing and methanogenic bacteria were quantified using a set of group specific probes (42, 43). Of particular note is the high abundance of *Desulfovibrio* sp. present both in the presence and absence of sulfate.

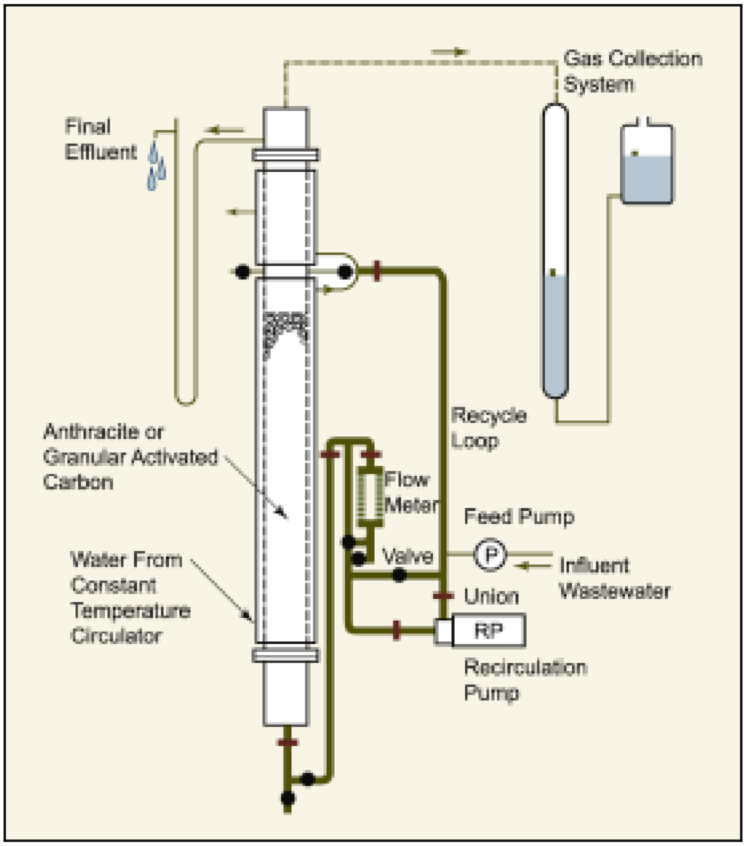


Figure 2.6. Schematic diagram of a fluidized (expanded) bed, anaerobic reactor (from 47).

Available data indicate that in the absence of sulfate, *Desulfovibrio* bacteria were growing in syntrophic association with methanogenic population(s) (37). Growth via syntrophic association is almost certainly a common state in the environment, and highlights the importance of evaluating stress response in a community context. For example, the physiological state of a *Desulfovibrio* species growing in syntrophic association with a methanogen could not be replicated in any pure culture scenario. We also anticipate that these reactors will be amenable to the introduction of the target organism (*Desulfovibrio vulgaris*), mutant variants, and closely related strains/species recovered from field sites and reactor systems established from field inocula. This aspect of the work will be done in close association with the Functional Genomics Core.

The fixed-bed reactor design is also suited to operation under very low organic loading, conditions comparable to many natural systems. For example, the fixed-bed reactor system (described above) was operated at a glucose feed (the sole carbon and energy source) of between 10–30 mg/L. These early studies also demonstrated excellent comparability between independent reactors. Replicate reactors demonstrated comparable population composition and reaction rates (e.g., sulfate removal and methane production), good resilience to change in operating conditions, and that the original community composition was reestablished following long-term perturbation (44). These general attributes—good experimental control, replication, and the maintenance of sufficient biomass for molecular characterization sites—are essential to the proposed studies.

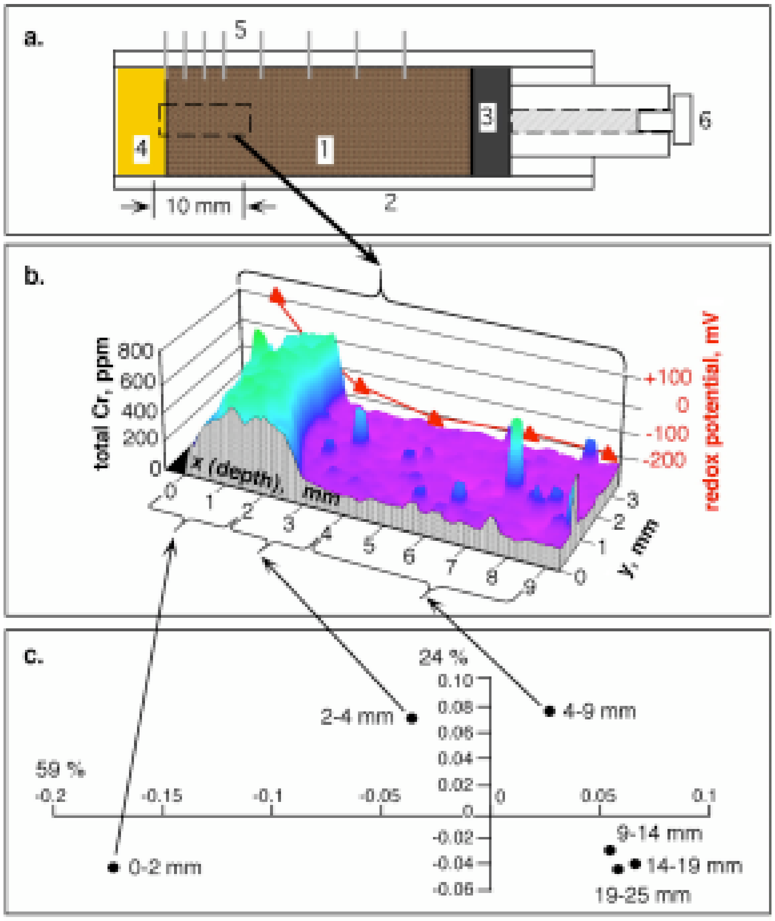


Figure 2.7. a. Synthetic soil aggregate column design. (1) Homogeneous soil pack, (2) plastic column, (3) piston for soil extrusion, (4) boundary reservoir for Cr(VI) solution, (5) Pt redox electrodes, and (6) plug. **b.** X-ray microprobe map of total Cr in a synthetic soil aggregate, **c.** Principal component plot of bacterial terminal restriction fragment length polymorphism (TRFLP) patterns from the synthetic soil aggregate show changes in bacterial community composition occurring in domains defined by [Cr] and redox potential (52).

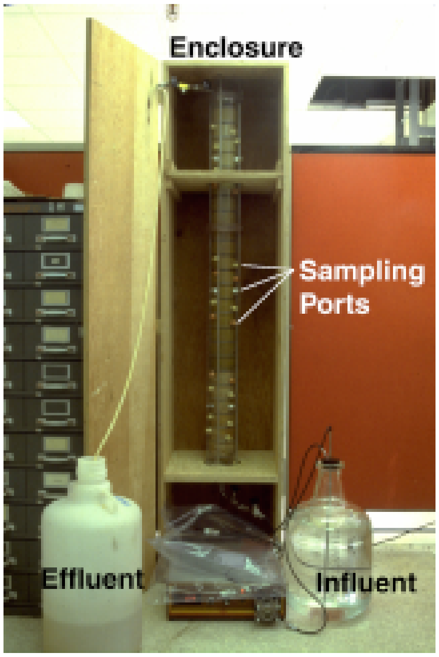


Figure 2.8. Large soil column for creating gradients of conditions typical of subsurface environments. Multiple sampling ports. Upflow system packed with sediment from study site. Can also be used in chemostat mode (10).

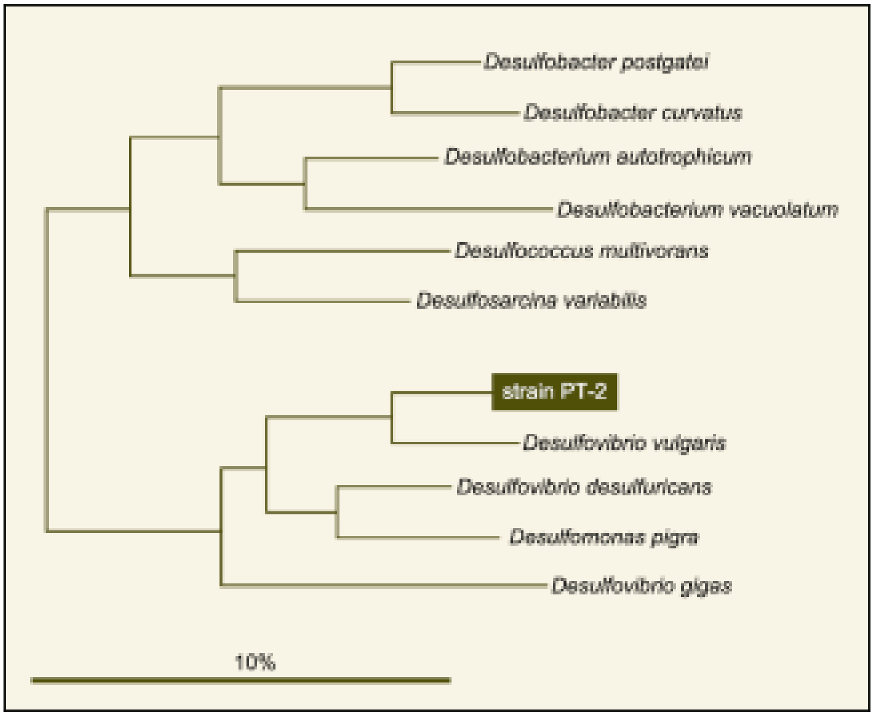

Desulfobacter postgatei
Desulfobacter curvatus
Desulfobacterium autotrophicum
Desulfobacterium vacuolatum
Desulfococcus multivorans
Desulfosarcina variabilis
strain PT-2
Desulfovibrio vulgaris
Desulfovibrio desulfuricans
Desulfomonas pigra
Desulfovibrio gigas
10%

Figure 2.9. Phylogenetic affiliation of bioreactor isolate PT2. This organism comprised a significant fraction of the *Desulfovibrio* population identified in the population structure study presented in Figure 2.5b. A pure culture was isolated with a fluorescent probe, developed using a 16S rRNA sequence recovered from reactor biomass, to screen enrichment cultures. We propose to further develop this general technique using combined FISH and high-speed sorting to recover specific populations related to our target organism for molecular characterization.

2.4.3.1. Rationale for Reactor Configuration. The general rationale for our selection of these reactor formats are summarized as follows:

- Analogous to saturated soil/sediment systems in which the microbiota colonize surfaces as biofilms
- Growth as biofilms is the more common state for microorganisms in the environment
- Fluidized bed reactor systems provide for experimental control, replication, high biomass, and ease of sampling
- Biofilm-associated populations are more resistant to stress and are the more relevant state for most environmental populations
- Biofilm-associated populations are not washed out (of the reactor system) following imposition of stress, providing for greater range of stress-response experimentation
- Past reactor studies have shown that *Desulfovibrio* spp.. and *Geobacter*-like populations will colonize and be retained within these systems

The fluidized bed reactor configuration (Figure 2.6) provides a more flexible format for sampling than the fixed bed and does not develop stratified structure as may occur with fixed bed reactors, even when operated at a high recycle rate (Figure 2.5a). The fluidized reactor will initially be set up with a sand bed, although other bed materials (e.g., small glass beads) can be examined in consideration of the initial characterization of population structure. The soil column format provides a colonization substratum that is more representative of the natural system, but also is prone to stratification and development of regions of heterogeneity, complicating sampling and analysis strategies. We have shown that while sampling is difficult, these heterogeneities can and need to be studied at a variety of scales (10, 52). As before, soil columns will be constructed at several scales from 10 µm to 1 m (Figures 2.5–2.8). Initially, we will use the smallest soil column simulations and gradually scale up to verify that our conceptual models of these systems' biogeochemistry are applicable to all scales. All reactors will be operated using a defined mineral salts/vitamin medium (described previously) amended with different organic substrates (e.g., glucose, acetate) (44).

2.4.3.2. Chemostat Studies. Although chemostats only poorly represent environmental conditions, they provide superb systems for controlled growth and comparative physiology. We will work closely with the Functional

Genomics Core on selected chemostat comparative analyses of the target organism and closely related organisms recovered from the field site and reactor system. Cultural isolation of populations closely related to *Desulfovibrio vulgaris* will be facilitated by the use of fluorescent probes designed to target 16S rRNA sequences recovered from environmental and reactor biomass as previously described (26).

2.4.3.3. Cultural Isolation of Organisms for Comparative "Benchmarks". As a more complete census of population structure in our field sites and reactor systems is developed using molecular techniques, we will initiate limited cultural isolation of populations related to our model organisms and those responding differentially to various imposed stresses (below). Although cultural isolation can be difficult, we are encouraged by our previous success in isolating a novel *Desulfovibrio* species from a fixed-bed reactor system (26). That isolation was facilitated by using a DNA probe designed on the basis of 16S rRNA sequence recovered directly from the reactor system to screen enrichments. We will work in close association with the other core groups to use these isolates as comparative "benchmarks" for identifying homologous stress response pathways, evaluating differential response to identified stressors, and for comparative studies of response in pure culture (chemostat, section 4.4.1) and in reactor systems from which they were isolated. Well-established methods will be used for isolation in culture (26).

2.4.3.4. Collection of Environmental Samples for Reactor Inoculation. Based on general chemical and microbiological mapping results, sediment/soil samples will be collected along the defined contamination/stressor gradients. Selected sampling locations will have similar natural characteristics to limit the environmental variables other than specific stressors contributing to microbial diversity. All sediment samples will be collected using clean "aseptic" drilling and handling techniques for microbiological specimens. Cores from anoxic areas will be put immediately in a nitrogen glovebox and either extracted on site to minimize "bottle effect" or transported under nitrogen and on ice to reduce any biotic or abiotic oxidation-induced changes. Depending on the stressor being studied samples may be shipped without cooling. All shipping will be overnight with a 24 h maximum hold time to ensure sample integrity.

2.4.3.5. Comparability of Reactor and Field Site Community Structure. The population structure of the different reactor systems in relationship to the field sites from which they were derived will be evaluated using a T-RFLP analysis similar to the one described above, but complemented by complete sequence analysis of 16S rRNA genes recovered using well-established methods of PCR amplification and cloning (30). This sequence information will be used to define general community architecture and to generate DNA probes targeting individual populations. Operating conditions will be varied to promote the development of communities representative of the field sites. Sequences will then be used to develop fluorescent probes for the selective recovery of specific populations using high-speed sorting (below) and cultural isolation (above).

2.4.4 Environmental Simulator Stress Analyses

2.4.4.1. Growth and Recovery of Stressed Single Populations under Simple Conditions. The first step in understanding the global molecular responses of *D. vulgaris* Hildenborough to stresses will be to establish the transcriptional and protein expression changes of the wild-type cells. To ensure reproducibility of the data, chemostat cultures of the SRB will be established at 30°C. The stresses to be examined initially are oxygen, heavy metal exposure (U, Co, Hg, Ni, and Cr), reductant limitation, phosphate restriction, and pH. Lactate will be the source of reductant with sulfate as the electron acceptor. Should the accumulated sulfide prove to be a serious problem in stabilizing a chemostat, alternative electron acceptors will be tested, such as fumarate or DMSO. *D. vulgaris* also ferments pyruvate, another condition in which sulfide can be minimized. Sulfidogenic cultures are likely to have significantly different responses to various heavy metals than cultures not producing sulfide. Both will be tested. All experiments will be carried out in triplicate. Stress conditions will be imposed by changes in the composition of the feedstock for the chemostat and will be followed over time in preliminary experiments to determine the appropriate sampling regimen. It is possible that the key regulatory signals will be rapid and transient.

To analyze the global molecular response of *S. oneidensis* MR-1 to stress, we will focus on elucidating the genetic basis of the cellular adaptation to changes in pH, phosphate availability, salinity, metal concentrations (e.g., Cd, Hg, Ni, and Cr), and nutrient availability (primarily carbon source). Since *S. oneidensis* is a facultatively anaerobic bacterium, we will not investigate the cellular response to high oxygen tension. Moreover, characterization of regulatory mechanisms governing the transition from aerobic to anaerobic growth is the primary objective of the MCP-funded project (the Shewanella Federation) aimed at studying the molecular basis of energy metabolism in *S. oneidensis*. Therefore, for this proposed GTL work, we will focus on the cellular response of *S. oneidensis* to pH, phosphate limitation, salinity, metal concentrations, and nutrient availability under both aerobic and anaerobic conditions. Experiments will be carried out using the cultivation protocols developed by the Shewanella Federation (Gorby and Fredrickson, unpublished), so that the microarray expression data generated by this project can be directly compared to those generated by relevant MCP- and MGP-funded projects.

Stress-response studies under anaerobic conditions will be conducted using fumarate as the electron acceptor. Due to the sensitive redox- and pH-dependent solubility of metals, studies of the physiological response of *S. oneidensis* cells to stress during metal reduction may be complicated by the formation of insoluble precipitates. Precipitate formation will especially become a factor when studying the effects of varying pH and phosphate and heavy-metal (U, Co, Hg, Ni, Cr) concentrations on global gene expression. Therefore, we will use fumarate as the model substrate to elucidate the cellular responses to pH, phosphate limitation, salinity, metal concentrations, and nutrient availability. The data generated by these experiments will be used to construct a stress-response model for *S. oneidensis*. However, since the long-term objective of the current proposal is to understand the effect of varying environmental conditions on bacterial metal reduction, the stress response model generated under fumarate reduction will be tested under Fe(III)-reducing conditions at later stages to determine the effects of salinity and varying carbon source on gene expression patterns.

For each condition, MR-1 cells will be grown in 7.5-L fermentors (New Brunswick Scientific) using defined M1 medium. Each experiment designed for evaluating gene expression under a given condition will be conducted using three biological replicates. For aerobic growth, the cultures will be supplemented with 20 mM lactate, while anaerobically grown *S. oneidensis* will be supplied with 20 mM lactate and 10 mM fumarate or 10 mM Fe(III)-nitrilotriacetic acid (NTA) as the electron acceptor. The MR-1 reference and test cells will be grown in continuous cultures under optimal physiological conditions at growth rates >90% Vmax with lactate as the limiting substrate. Upon reaching the steady-state conditions (synchronization), the test cultures will be subjected to a single stress factor, while reference cultures will continue to be grown under normal (nonstress) conditions. Samples from both reference and test cultures will be taken at 15 min increments up to 2 hr poststress. The cells will be harvested by brief centrifugation (10,000 × g for 3 min at 4°C).

In addition to studies of the target organisms, a selected set from the cultural isolates will also be characterized in the chemostats—i.e., it will be important to evaluate the growth of closely related, as well as more distantly related, organisms grown under similar experimental conditions to determine if observations made for the model organisms can be generalized.

2.4.4.2. Identification and Recovery of Stress-Responsive Populations from Complex Communities. The reactor communities will be transiently exposed to general and specific agents of stress as described in the previous section for single populations under simple conditions. The stressors to be considered are oxygen, pH, high nitrates, heavy metal exposure (U, Co, Hg, Ni, and Cr), reductant limitation, phosphate restriction, and salinity; lactate and fumarate will be used as electron donors and sulfate as the terminal electron acceptor under anaerobic conditions.

Since environments are not static but constantly changing with respect to key physical and chemical variables (e.g., substrate concentration, temperature, pH, pressure, salinity, osmolarity, light, redox potential, etc.), it follows that all populations in a natural system are not simultaneously growing optimally. Those populations growing under suboptimal conditions, or entering resistant (e.g., spores) or moribund states, are experiencing stress. Thus, we suggest that it is not sufficient to monitor an extrinsic parameter known to be associated with stress—it is essential to understand the environmental context of stress for individual populations. We hypothesize that within any well-adapted and dynamic system, some fraction of the community will be experiencing stress. This component of the community will be identified using a general activity measure (rRNA synthesis) to identify populations that respond to changing chemical or physical environments by entering a nongrowth, or low growth, state. Cells comprising specific stress-responsive populations will then be physically recovered using a combination of a high-speed cell sorter and fluorescent in situ hybridization (FISH), using fluorescent probes designed to target specific rRNA sequence types. In this way, we will achieve full integration of system elements, linking specific stress-regulated genes to specific populations shown to respond differentially to specific stresses.

Note: These studies will be directly integrated and in support of the expression analysis of the Functional Genomics Group (Subsection 3.4.2.2) and modeling studies; (Subsections 4.3 and 4.4). For example, these systems will later be used to monitor whole-system stress response using microarrays developed from the collection of stress responsive genes and related to model predictions, etc.

2.4.4.3. Radiomicroarray Analysis. Change in growth status will be evaluated using a radiomicroarray format to measure changes in the rRNA synthesis of individual populations resident in the reactor systems described above. The microarray will build upon our (Stahl) research program developing a platform for comprehensive monitoring of microbial populations in complex communities. Unlike the more standard glass-slide array, the gel-based format immobilizes DNA probes in small acrylamide pads (100 □ 100 □ 20□m) arrayed on a glass surface. This provides for much higher probe density (and target capture capacity) than possible by direct immobilization on the glass surface. Probes immobilized on individual pads are designed to complement specific rRNA sequence types, either unique to individual species or encompassing larger phylogenetic groups. Our studies have demonstrated the utility of this format to directly detect environmental rRNAs using an optimized fragmentation and fluorescent dye labeling protocol (15). This microarray format has been well characterized with respect to specificity and reproducibility (15, 31, 53).

We have recently combined the gel-based array with radiolabeling (Figure 2.10). In this format the microbial community is labeled by adding ^{33}P phosphate (or ^{3}H uridine) to the growth medium immediately following perturbation/stress (e.g., altered pH, metal addition). The radioactive phosphate is incorporated only into the nucleic acids of actively growing populations. Following recovery of total rRNA and hybridization to the microarray, the array is coated with a silver halide emulsion. The emulsion is developed following an appropriate period of incubation, and radioactivity of the rRNA bound to each individual gel pad is quantified by observing precipitated silver grains with a diameter of approximately 0.2 µm using microscopy. Thus, nongrowing populations are labeled only with fluorescent dye, whereas active populations are identified by generating both fluorescent dye and silver grain deposition signals. We recognize that not all cells of a specific population will be of comparable physiological status, for example, if distributed at different depths in the biofilm. However, this analysis format using replicate reactor systems will measure average shifts in activity (growth) profiles.

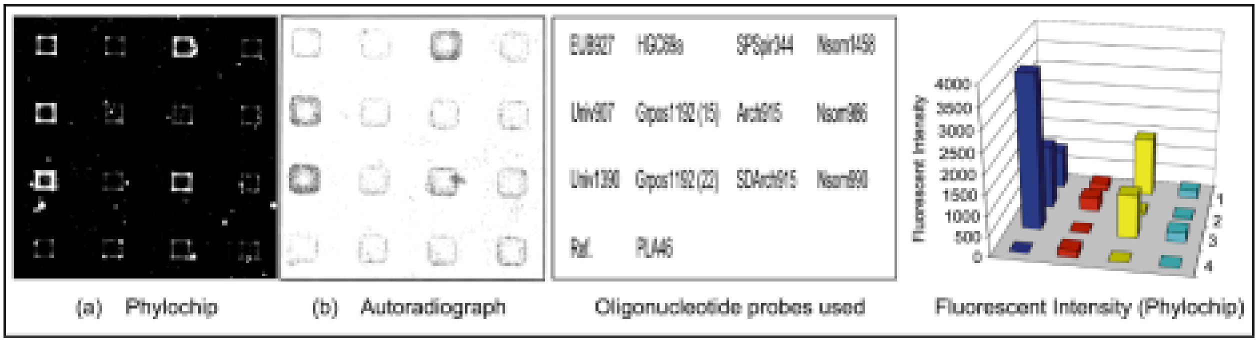


Figure 2.10. Schematic of radiomicroarray method. (a) The *Phylochip* microarray consists of 2,704 gel pads (100 □ 100 □ 20 □m) arrayed in four quadrants on a microscope slide. Each pad contains 1–10 pmole of DNA probes, which are complementary to the rRNA of a specific species or phylogenetic group (~ ranks of genus, family, and higher taxons). The general probes on the microarray target domains (bacteria, archaea), subdivisions within the proteobacteria, and different genera of gram-positive bacteria. The hybridization pattern (right panel) was obtained using native environmental rRNA recovered from less than 1 g of marine sediment (El Fantroussi, et al., in preparation). (b) Radiomicroarray. After the phylochip slide format was read fluorescently, the slide was dipped in photographic emulsion and exposed. Silver grains were formed over gel pads that have annealed radioactive target rRNA.

2.4.4.4 Recovery of Stress Responsive Populations via Flow Cytometry. Flow cytometry is a well-established method of measuring physical characteristics of individual cells, including light scatter (size and shape) and fluorescence emission at wavelengths of interest. This technique was initially used in biological oceanography to quantify and sort naturally fluorescent marine picoplankton. Our earlier work documented the feasibility of using flow cytometry in combination with FISH to quantify specific microbial populations using probe-conferred fluorescence (1) and this method has now received general use in environmental microbiology (13).

Our more recent studies have shown that high-speed sorting can be used to physically recover specific populations, using FISH as the sorting parameter. Figure 2.11 shows the results of a successful sort in which a specific population of cells was recovered based on probe-conferred fluorescence. The pooled cells then serve for general and specific genetic characterization. For example, a few thousand sorted cells were sufficient to recover 16S rRNA genes using PCR amplification. We anticipate that sufficient numbers of cells can be recovered for use in other sequence-based analyses, including hybridization and large-fragment cloning for the directed recovery of homologous stress response genes.

The proposed combination of methods provides a technical framework to link a physiological response of a specific population within a community setting (radiomicroarray) and to associate that population with a specific genetic system of stress response. We will continue to work closely with Dr. van den Engh and staff at the Institute of Systems Biology, Seattle (ISB), on using and optimizing flow cytometry for the proposed research.

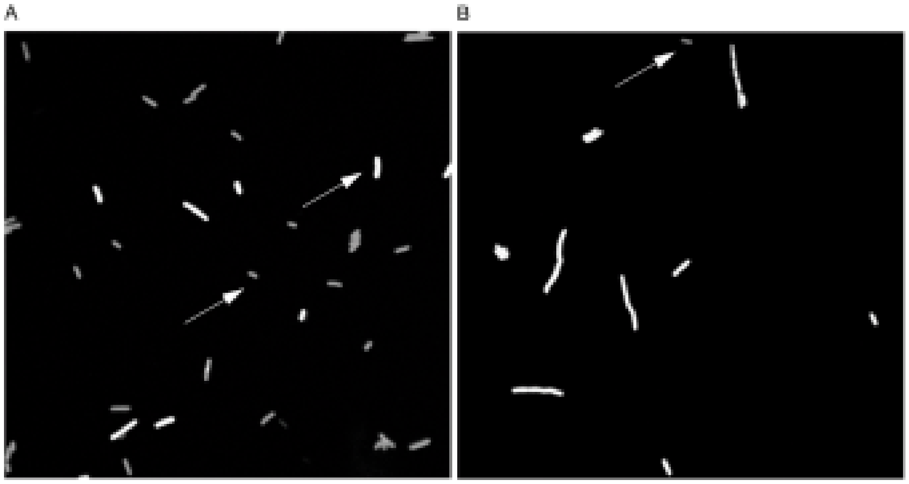


Figure 2.11. Fluorescence photomicrograph of bacterial cells fixed with paraformaldehyde and labeled by hybridization with fluorescently labeled, group-specific 16S rRNA oligonucleotide probes. Panel A shows a mixture of *Escherichia coli* and *Pseudomonas aeruginosa* (noted with arrows) cells before flow cytometric sorting. Panel B shows the same population of cells after sorting. The arrow denotes a single cell not properly excluded during the sort. Courtesy of A. Schramm.

2.4.4.5. Flow Cytometric Sorting of FISH-labeled Populations. Flow cytometry has been productively used to provide general information of single-cell abundance and certain aspects of their activities (e.g., cell size, DNA and RNA content, membrane potential). However, most past applications have been of relatively coarse resolution because analyses were restricted to populations having unique intrinsic properties such as autofluorescence (7, 38).

The use of fluorescent dye-labeled probes to label single cells via fluorescence in situ hybridization (FISH) now provides the basis to selectively confer fluorescence on any population using genetic criteria (Figure 2.12). Our group contributed to both the early development of the basic format and the early documentation of the combined use of FISH and flow cytometry (1). A key goal of our proposed research will be to use fluorescent probes complementary to the 16S rRNAs to selectively recover specific environmental and reactor populations via flow cytometric sorting, providing a mechanism to associate genes recovered from the bulk community with specific members of the community. Probes will be developed to target specific populations highlighted by the radiomicroarray studies. The success of sorting a population from reactor systems and field-site samples will in part depend on our ability to isolate a relatively clean cellular fraction from a reactor or environmental substratum. We will evaluate published protocols for isolating cells from surfaces and sediments, using sonication combined with differential and density gradient centrifugation (3). We will also evaluate the use of enrichment cultures developed from reactor effluent, eliminating or reducing the need for fractionation of cells from a solid substratum. Sorted populations will be further characterized using PCR and hybridization to link them with specific genes implicated in stress response pathways. We will work in close association with Dr. Ger van den Engh (of ISB), who has a well-established program in the development and application of high-speed sorters (see letter of collaborative support.

2.4.4.6. Expression Analysis of RNA Extraction from Sediment columns and Reactor Systems. Environmental simulators play a critical role as source material for expression analysis stress response genes in a community setting. These analyses will target specific populations naturally resident (e.g., *Desulfovibiro* spp.) or introduced into these systems (e.g., mutant strains). The following is an example of molecular analyses that will be performed on soil samples from the reactors and the environmental samples. Soil samples will be mixed with 2 g of sterile sand and transferred to a sterile mortar sand (51). Samples will be mixed with 1 ml denaturing solution [4 M guanidine thiocyanate, 50 mM b-mercaptoethanol, 10 mM Tris-HCL (pH 7.0), 1 mM EDTA, 0.5% sarsokosyl],

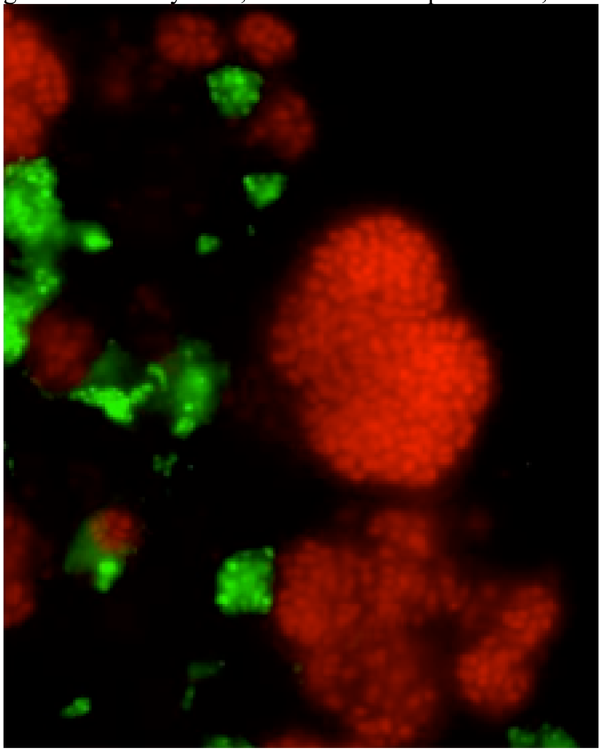


Figure 2.12. FISH micrograph, *Nitrobacter* sp using nitrite oxidizers (36).

frozen in liquid nitrogen, and ground until thawed (24). The process of freezing and grinding will be repeated twice. After the addition of 9 ml of extraction buffer [100 mM sodium phosphate (pH 7.0), 100 mM Tris-HCL (pH 7.0), 100 mM EDTA (pH 8.0), 1.5 M NaCl, 1% hexadecyltrimethylammonium bromide, and 2% SDS], the samples will be incubated for 30 min at 65°C with occasional gentle mixing, and centrifuged at 1,800 □ g for 10 min. The supernatants will be collected into prechilled tubes that contain 20 ml 24:1 chloroform:isoamyl alcohol. The nucleic acid in the aqueous phase will be precipitated with 0.6 volume of isopropyl alcohol for 30 min at room temperature, and centrifuged at 16,000 □ g for 20 min at 20°C. The crude nucleic acid pellet will then be suspended in 1 ml of DEPC-treated D_2O and passed through 5 ml of a Sephadex G-75 resin slurry (Pharmacia) to remove organic contaminants. The RNA will be purified from the nucleic acid with a Qiagen Tip 20, and treated with RQ1 RNase-free DNase I (Promega) according to the manufacturer’s instructions (see Section 3.4.2) [expression array technology]).

2.4.4.7 Synchrotron FTIR Direct Analysis of Stress Changes in Living Cells. The ability to image and characterize microbial responses to stresses in complex and extremely inhomogeneous and transient natural environments has been one of the most challenging areas. Recent work from the Holman group (21-23) suggests that one can meet this challenge by using synchrotron-radiation-based Fourier transform infrared (SR-FTIR) spectromicroscopy, especially by combining SR-FTIR results with information from other bioanalytical and imaging techniques. Here, we propose to use this as a real-time technique to characterize microbial resistance to stress.

Synchrotron radiation is utilized because the infrared photons extracted from the radiation source can be focused to a spot less than 10 □m(in diameter). The LBNL Advanced Light Source, the brightest radiation source in the world, routinely obtains high infrared intensities with a signal-to-noise ratio that is at least 100 times better than any infrared thermal source at a spatial resolution of 3 to 10 □m (33). This unique property allows very sensitive discrimination of IR spectra, even on geological materials that inherently have very low infrared reflective surfaces and at the same time often have a spatial variation of typically 3 to 10 □m. More recently, Holman’s group has proved experimentally that although the infrared light from ALS’s sychrotron radiation source is very intense, it neither heats up the sample (35) nor affects the physiology of a living biological system (Holman et al., in press). Unlike other surface chemistry microprobe techniques (e.g., UV or x-ray microprobe), the SR-FTIR spectromicroprobe is truly a nondestructive and noninvasive technique that provides for direct interrogation of the same location of geological materials with microorganisms. This provides researchers with an extremely beneficial opportunity to quickly track and image the fine-scale (3–10 □m) chemical changes in different compartments of a living biofilm on surfaces of geological materials without the need for staining procedures or chemical tags.

The significance of this monitoring and assessment technology breakthrough for studying microorganisms in geological environments was featured in *Chemical and Engineering News* (59) and *Optics and Photonic News* (27). It was also acknowledged at the American Physical Society Press Conference in 2000.

The environmental simulators and chemostats will be modified so that rock chips from the sampling site can be inserted into and removed from the chambers and reactors at different points to monitor biofilms on the surface using the SR-FTIR technique. This way, the resistance to stress in these systems can be measured, IR signature changes can be documented, and population changes can be imaged in the biofilm on the chip, all in real-time. This alternate method for measuring stress changes in the cell can also cross-verify the expression analysis studies for the stress regulation pathways in the environmental simulators and chemostats.

2.4.4.8. Direct Stress and Community Comparison with Phospholipid Fatty Acid (PLFA) and Metabolite Analyses. Phospholipid fatty acid analysis (PLFA) has been used extensively for characterizing and monitoring microorganisms in all types of environments. Most generally, this method has been used for biomass estimates. However, some PLFAs (signature compounds) are unique to a species or strain, or conserved across broad physiological groups, families, or even kingdoms. For example, PLFA signatures have been used to determine the environmental distribution of methanotrophs, actinomycetes, and anaerobes (17, 18, 41, 46). Certain groups of fatty acids (cis and trans isomers) are also known to change in response to altered physiological status, and these have been used to monitor the status of environmental populations. The normal detection limit is about 10,000 cells; however, recent studies have shown that much lower densities can be detected using microassay techniques (17).

The extraction process for the lipids has also been used to simultaneously extract nucleic acids, providing a method to relate sequence information (e.g., as determined by DNA probing) and lipid signature compounds from the same environmental sample.

We will use this approach to relate sequence-specific information (e.g., T-RFLP mapping, DNA probing) to lipid profiles in actinide-rich environments. The combined analyses will provide a way to associate populations identified by sequence with molecular signatures associated with physiological status. Water will be filtered in situ or in the field, and the filters frozen (–70°C) for later assay. Sediment and groundwater samples will be stored at –70°C for later laboratory assay. We will conduct similar studies using our controlled bioreactor systems.

Labeled lactate (^{13}C) will also be used in combination with the PLFA analyses to determine the amount of lactate that is being incorporated into the extant biomass and which groups of the microbial community that are using it under stress simulation conditions. This will be done by measuring the isotope ratios of ^{13}C:^{12}C in specific signature compounds as well as the amount of ^{13}C in the total lipids. We have substantial experience with the combination of PLFA and stable isotope analysis using both ^{13}C-enriched substrates and natural abundance levels of ^{13}C (57). Because the lactate will be substantially enriched in ^{13}C, the small amount of naturally occurring isotopic discrimination that occurs during synthesis of fatty acids is not an issue. These isotope ratio mass spectrometer techniques have been developed in collaboration with Dr. Mark Conrad at LBNL. Location and quantification of microbial activities in the environmental simulators will also be accomplished indirectly by analysis of small (20 µl) subsamples from the soil columns and reactor systems for daughter products and metabolites. For example, NO_{3-} disappearance will be assayed as an indicator of denitrification and TOC reduction for general heterotrophic activity. Our standard assay procedure for NO_{3-} (Lachat automated flow colorimetry) provides more than adequate sensitivity; our TIC-TOC analyzer for TOC may have sufficient sensitivity for the higher concentrations of lactate.

2.4.5 Validation of Conceptual Models with Environmental Simulators and Field Tests

By the final two years of this project, we expect that the computational group will have developed a conceptual model of stress regulatory pathways. These models will be tested by subjecting various reactor systems to various stress and recovery scenarios and will be analyzed with the parameter presented in Subsection 2.4.4. We also anticipate that we will have developed a number of new molecular probes that we can use to verify the responses and pathways being used for various stressors. Once we have validated these conceptual models in our controlled laboratory simulations, we will use the same probes to validate these conceptual models in the field. The field tests will use the same techniques described for the original survey and mapping studies with the addition of the newly developed stressor pathway molecular probes. Field tests will involve push-pull tests similar to those already performed at the NABIR FRC, along defined stressor gradients (25). Depending on the stressor being tested amendments in the form of electron donors, electron acceptors, nutrients, etc., may be added to the environment to measure the expected stressor change. These field tests will be developed in detail in consideration of the test site, e.g., in association with the NABIR FRC Manager. Once the conceptual models have been developed, a number of other field tests might also be considered at other DOE sites. Questions to then ask could include: (1) are *Desulfovibrio*-like, *Shewanella*-like, and *Geobacter*-like organisms found at other contaminated sites; (2) are the metabolites, genes, and pathways present at most sites; (3) is there a direct correlation between the presence of certain genes and pathways at sites with certain types and concentrations of contaminants; and (4) can we induce particular stress responses that might favor a desired outcome for contaminants in the subsurface at some sites.

2.5 Experimental Core Facilities

The Environmental Molecular Microbiology Facility (ORNL, LBNL, U. Washington, Diversa Inc.), Environmental Simulation and Culture Facility (LBNL, U. Washington, Diversa Inc.). Technologies and facility resources are under development at LBNL and Diversa for culturing hard to grow organisms in controlled, reproducible environments. LBNL is also developing defensible environmental model chambers at different scales for better mimicking the natural environment of the microorganisms under study. These facilities will be fully integrated with the Advanced Light Source Microscope Beam Lines (1.4.3, 4.0.1-2, 6.1.2, 7.0.1, 10.3.1, 10.3.2) to take advantage of it's unique analytical capabilities for environmental and biological samples, ie. infra red and x-rays; the Center for Isotope Geochemistry and it's ability to analyse environmental samples for stable isotopes, eg. isoprobe; and the Center for Environmental Biotechnology with it's facilities for PLFA and nucleic acid analyses from environmental samples, soil columns, bioreactors, SLCM imaging, and biological safety level 2 laboratory. The LBNL National Center for Electron Microscopy will also enable other studies of environmental and biological specimens using state-of-the-science electron microscopes, eg. scanning transmission electron microscopy. This integration of unique instrumentation and facilities at LBNL with the VIMSS facilities at University of Washington for flow cytometry, bioreactors, and functional microarray construction; and the Diversa environmental culture, isolation, archiving facilities; make this facility a unique and valuable resource for DOE. These facilities are also key to obtaining the highest quality and quantity of biological material for the other experimental facilities and research cores in VIMSS.

2.6 References

1. Amann, R. I., B. Binder, S. W. Chisholm, R. Olsen, R. Devereux, and D. A. Stahl. Combination of phylogenetically based fluorescent hybridization probes and flow cytometry. 1990. *Appl. Environ. Microbiol.* **56**:1619–1625.
2. Angers, D. A., and M. Giroux. 1996. Recently deposited organic matter in soil water-stable aggregates. Soil Sci. Soc. Am. J. **60**:1547–1551.
3. Bakken, L. R. 1985. Separation and Purification of Bacteria from Soil. Appl. Environ. Microbiol. **49**:1482–1487.
4. Braker, G., H. L. Ayala-del-Rio, A. H. Devol, A. Fesefeldt, and J. M. Tiedje. 2001. Community structure of denitrifiers, Bacteria, and Archaea along redox gradients in pacific northwest marine sediments by terminal restriction fragment length polymorphism analysis of amplified nitrite reductase (nirS) and 16S rRNA genes. Appl. Environ. Microbiol. **67**:1893–190.
5. Britschgi, T. B., and S. J. Giovannoni. 1991. Phylogenetic analysis of a natural marine bacterioplankton population by rRNA gene cloning and sequencing. Appl. Environ. Microbiol. **57**:1707–1713.
6. Burn, Y. V., and L. Shapiro. 1992. A temporally controlled sigma-factor is required for polar mophogenesis and normal cell division in *Caulobacter*. Genes. Dev. **6**:2395–2408.
7. Button, D. K., and B. R. Robertson. 2001. Determination of DNA content of aquatic bacteria by flow cytometry. Appl. Environ. Microbiol. **67**:1636–1645.
8. Currie, J. A. 1961. Gaseous diffusion in the aeration of aggregated soils. Soil Sci. **92**:40–45.
9. Dilling, W., and H. Cypionka. 1990. Aerobic respiration in sulfate-reducing bacteria. FEMS Microbiol. Lett. **71**:123–128.
10. Enzien, M. V., F. Picardal, T. C. Hazen, R. G. Arnold, and C. B. Fliermans. 1994. Reductive dechlorination of trichloroethylene and tetrachloroethylene under aerobic conditions in a sediment column. Appl. Environ. Microbiol. **60**:2200–2205.
11. Franzluebbers, A.J., and M.A. Arshad. 1997. Soil microbial biomass and mineralizable carbon of water-stable aggregates. Soil Sci. Soc. Am. J. **61**:1090–1097.
12. Fry, N. K., J. K. Fredrickson, S. Fishbain, M. Wagner, and D. A. Stahl. 1997. Population structure of microbial communities associated with two deep, anaerobic, alkaline aquifers. Appl. Environ. Microbiol. **63**:1498–1504.
13. Fuchs, B. M., G. Wallner, W. Beisker, I. Schwippl, W. Ludwig, and R. Amann. 1998. Flow cytometric analysis of the in situ accessibility of Escherichia coli 16S rRNA for fluorescently labeled oligonucleotide probes. Appl. Environ. Microbiol. **64**:4973–4982.
14. Gorby, Y. A., and D. R. Lovley. 1992. Enzymatic uranium precipitation. Environ. Sci. Technol. **26**:205–207.

15. Guschin, D. Y., B. K. Mobarry, D. Proudnikov, D. A. Stahl, B. E. Rittmann, and A. D. Mirzabekov. 1997. Oligonucleotide microchips as genosensors for determinative and environmental studies in microbiology. Appl. Environ. Microbiol. **63**:2397–2402.
16. Hardy, J. A., and W. A. Hamilton. 1981. The oxygen tolerance of sulfate-reducing bacteria isolated from North Sea waters. Curr. Microbiol. **6**:259–262
17. Hazen, T. C. 1997. Bioremediation. In: Microbiology of the Terrestrial Subsurface (eds) P. Amy and D. Haldeman, p 247–266. CRC Press, Boca Raton.
18. Heipieper, H. J., B. Loffeld, H. Keweloh, and J. A. M. Debont. 1995. The cis/trans isomerization of unsaturated fatty-acids in Pseudomonas putida s12 - an indicator for environmental-stress due to organic-compounds. Chemosphere **30**:1041–1051.
19. Holman, H.-Y.N., K. A. Bjornstad, M. P. McNamara, M. C. Martin, W. R. McKinney, E. A. Blakely. *In press.* Journal of Biomedical Optics.
20. Holman, H.Y.N., Goth-Goldstein, R. Blakely, E.A., Martin, M.C., McKinney, W.R. Abstracts of Papers American Chemical Society 2000, 219, ANYL 105.
21. Holman, H.-Y.N., Goth-Goldstein, R., Martin, M.C., Russell, M.L., McKinney, W.R. 2000. Environ. Sci. Tech. **34**:2513–2517.
22. Holman, H.-Y.N.; Nieman, K.; Sorensen, D.L.; Miller, C.D.; Martin, M.C.; Borch, T.; McKinney, W.R.; Sims, R.C. 2002. Environ. Sci. Tech. **36**:1276–1280.
23. Holman, H.-Y.N.; Perry, D.L.; Martin, M.C.; Lamble, G.M.; McKinney, W.R.; Hunter-Cevera, J.C. 1999. Geomicrobiology J. **16**:307–324.
24. Hurt, R. A., X. Qiu, L. Wu, Y. Roh, A.V. Palumbo, J. M. Tiedje, and J. Zhou. 2001. Simultaneous recovery of RNA and DNA from soils and sediments. Appl. Environ. Microbiol. 67:4495–4503.
25. Istok, J.D., Humphrey, M.D., Schroth, M.H. Hyman, M.R., and O'Reilly, K.T., 1997. single-well, "push-pull" test method for in situ determination of microbial metabolic activities. Ground Water **35**(4): 619–631.
26. Kane, M. D., L. K. Poulsen, and D. A. Stahl. 1993. Monitoring the enrichment and isolation of sulfate-reducing bacteria by using oligonucleotide probes designed from environmentally-derived 16S rRNA sequences. Appl. Environ. Microbiol. **59**:682–686.
27. Krupa, T. 1999. Infrared Beamline Reveals Toxin-Reducing Microbes. Optics & Photonics News **10**(11):8.
28. Leffelaar, P. A. 1986. Dynamics of partial anaerobiosis, denitrification, and water in a soil aggregate: Experimental. Soil Sci. **142**:352–366.
29. Liu, W-T. and D. A. Stahl. 2001. Molecular approaches for the measurement of density, diversity, and phylogeny. *In:* Manual of Environmental Microbiology, Second Edition. ASM Press, Washington, D.C. pp. 114–134.
30. Liu, W-T., A.D. Mirzabekov, and D.A. Stahl. 2001. Optimization of an Oligonucleotide Microchip for Microbial Community Structure Studies: A Non-Equilibrium Dissociation Approach. Environ. Microbiol. **3**:619–629.
31. Liu, W.-T., T. L. Marsh, and L. J. Forney. 1997. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of 16S ribosomal DNA. Appl. Environ. Microbiol. **63**:4516–4522.
32. Lovley, D. R., P. K. Widman, J. C. Woodward, and E. J. P. Phillips. 1993. Reduction of uranium by cytochrome c3 of Desulfovibrio vulgaris. Appl. Environ. Microbiol. **59**:3572–3576.
33. Lumppio, H. L., V. Shenvi Neeta, A. O. Summers, G. Voordouw, and D. M. Kurtz, Jr. 2001. Rubrerythrin and rubredoxin oxidoreductase in Desulfovibrio vulgaris: A novel oxidative stress protection system. J. Bacteriol. **183**:101–108.
34. Martin, M.C.; McKinney, W.R. 2001. Ferroelectrics **249**:1–10.
35. Martin, M.C.; Tsvetkova, N.M.; Crowe, J.H.; McKinney, W.R. 2001. Appl Spectrosc **55**:111–113.
36. Mobarry, B.K., M. Wagner, V. Urbain, B.E. Rittmann, and D.A. Stahl. 1996. Phylogenetic probes for analyzing abundance and spatial organization of nitrifying bacteria. Appl. Environ. Microbiol. **62**:2156–2162.

37. Noguera, D. R., G. A. Brusseau, B. E. Rittmann, and D. A. Stahl. 1998. A unified model describing the role of hydrogen in the growth of *Desulfovibrio vulgaris* under different environmental conditions. Bioeng. Biotechnol. **59**:733–746.

38. Olson, R. J., S. W. Chisholm, E. R. Zettler, and E. V. Armbrust. 1988. Analysis of *Synechococcus* pigment types in the sea using single and dual beam flow cytometry. Deep-Sea Res. **35**:425–440.

39. Peck, Jr., H. D., and J. LeGall (eds). 1994. Inorganic microbial sulfur metabolism. Meth. Enzymol. Vol. 243.

40. Postgate, J. R. 1984. The sulphate reducing bacteria, ed. 2. Cambridge University Press.

41. Phelps, T. J., D. Ringelberg, D. Hedrick, J. Davis, C. B. Fliermans, and D. C. White. 1989. Microbial biomass and activities associated with subsurface environments contaminated with chlorinated hydrocarbons. Geomicrobiol J. **6**:157–170.

42. Raskin, L., L. K. Poulsen, D. R. Noguera, B. E. Rittmann, and D. A. Stahl. 1994. Quantification of methanogenic groups in anaerobic biological reactors using oligonucleotide probe hybridizations. Appl. Environ. Microbiol. **60**:1241–1248.

43. Raskin, L., W. C. Capman, M. D. Kane, B. E. Rittmann, and D. A. Stahl. 1996. Critical evaluation of membrane supports for use in quantitative hybridizations. Appl. Environ. Microbiol. **62**:300–303

44. Raskin, L., B. E. Rittmann, and D. A. Stahl. 1996. Competition and co-existence of sulfate-reducing and methanogenic populations in anaerobic biofilms. Appl. Environ. Microbiol. **62**:3847–3857.

45. Raskin, L., J. M. Stromley, B. E. Rittmann, and D. A. Stahl. 1994. Group-specific 16S rRNA hybridization probes to describe natural communities of methanogens. Appl. Environ. Microbiol. **60**:1232–1240.

46. Ringelberg, D. B., G. T. Townsend, K. A. Deweerd, J. M. Suflita, and D. C. White. 1994. Detection of the anaerobic dechlorinating microorganism Desulfomonile tiedjei in environmental matrices by its signature lipopolysaccharide branched-long-chain hydroxy fatty-acids. Fems Microbiology Ecology **14**:9–18.

47. Rittmann, B. E., L. A. Crawford, C. K. Tuck, and E. Namkkung. 1986. *In Situ* determination of kineteic parameters for biofilms: Isolation and characterization of oligotrophic biofilms. Biotechnol. Bioeng. **28**:1753–1760

48. Smith, K. A. 1977. Soil Aeration. Soil Sci. **123**:284–291.

49. Stahl, D. A. and R. Amann. 1991. Development and application of nucleic acid probes in bacterial systematics, p. 205–248. *In* E. Stackebrandt and M. Goodfellow (ed.), Sequencing and Hybridization Techniques in Bacterial Systematics. John Wiley and Sons, Chichester, England.

50. Tokunaga, T. K., S. R. Sutton, and S. Bajt. 1994. Mapping of selenium concentrations in soil aggregates with synchrotron x-ray fluorescence microprobe. Soil Sci. **158**:421–433.

51. Tokunaga, T.K., S.R. Sutton, S. Bajt, P. Nuessle, and G. Shea-McCarthy. 1998. Selenium diffusion and reduction at the water-sediment boundary: Micro-XANES spectroscopy of reactive transport. Environ. Sci. Technol. **32**:1092–1098.

52. Tokunaga, T. K., J. Wan, M. K. Firestone, T. C. Hazen, E. Schwartz, S. R. Sutton, M. Newville. 2001. Chromium diffusion and reduction in soil aggregates. Environ. Sci. Technol. **35**:3169–3174.

53. Urakawa, H., P. A. Noble, S. El Fantroussi, J. J. Kelly, and D. A. Stahl. 2002. Single-Base Pair Discrimination of Terminal Mismatches Using DNA Microarrays and Neural Network Analyses. Appl. Environ. Microbiol. **68**:235–234.

54. Urakawa, H., T. Yoshida, M. Nishimura, and K. Ohwada. 2000. Characterization of depth-related population variation in microbial communities of a coastal marine sediment using 16S rDNA-based approaches and quinone profiling. Environmental Microbiology **2**:542–554.

55. U.S. Department of Energy, Description of the Bear Creek Valley Field Research Site, http://www.esd.ornl.gov/BCV-FieldSite/

56. Voordouw, G. 1993. Molecular biology of the sulfate-reducing bacteria. In: The Sulfate-Reducing Bacteria: Contemporary Perspectives. Odom, J.M., Singleton, R., Jr., eds. Springer-Verlag, New York pp. 88–130.

57. Waldrop, M.P., T.C. Balser, and M.K. Firestone. 2000. Linking microbial community composition and function in a tropical soil. Soil Biol. Biochem. **32**:1837–1846.

58. Wang, Y.-T., M.T. Suidan, and B.E. Rittmann. 1985. Performance of expanded-bed methanogenic reactor. J. Environ. Eng. **111**:460–471.

59. Watson, D. B., B. Gu, D. H. Phillips, and S. Y. Lee. 1999. Evaluation of Permeable Reactive Barriers for Removal of Uranium and other Inorganics at the Department of Energy Y-12 Plant, S-3 Disposal Ponds. ORNL/TM-1999/143, Oak Ridge National Laboratory, Oak Ridge, TN.

60. Yan, T., M.W. Fields, S.L. Carroll, J.M. Tiedje, and J. Zhou. 2002. The diversity of nitrite reductase genes (nirK and nirS) from bacterial communities and isolates in nitrate- and uranium-contaminated groundwater. American Society for Microbiology 102nd General Meeting, Salt Lake City, UT.

61. Zausig, J., W. Stepniewski, and R. Horn. 1993. Oxygen concentration and redox potential gradients in unsaturated model soil aggregates. Soil Sci. Soc. Am. J. **57**:908–916,1993.

62. Zhou, J., M.A. Burns and J.M. Tiedje. 1996. DNA recovery from soils of diverse composition. Appl. Environ. Microbiol. **62**:316–322.

3 FUNCTIONAL GENOMICS

Jay Keasling (LNBL/UCB, Core Team Leader)
Alex Beliaeve (ORNL)
Carolyn Bertozzi (LBNL/UCB/HHMI)
Matthew Fields (ORNL)
Martin Keller (DIVERSA Inc.)
Anup Singh (SNL)
Dorthea Thompson (ORNLO)
Judy Wall (University of Missouri)
Jizhong Zhou (ORNL)

3.1 Goals and Specific Aims

The goal of the Functional Genomics Core is to develop the experimental methods to elucidate the regulatory networks in the stress responses of *Desulfovibrio vulgaris*, *Shewanella oneidensis*, and *Geobacter metallireducens.* Specifically, we propose to

1. use existing DNA arrays for *S. oneidensis,* and to develop DNA arrays for *D. vulgaris* and *G. metallireducens* to measure the transcript profile (transcriptome) during the response to various environmental stresses;
2. use HPLC-MS-MS to measure the protein profile (proteome) of *D. vulgaris*, *G. metallireducens*, and *S. oneidensis* during the response to various environmental stresses;
3. measure the metabolite profile (metabolome) of *D. vulgaris*, *G. metallireducens*, and *S. oneidensis* during the response to various environmental stresses;
4. determine protein-protein interactions in the signaling cascade of *D. vulgaris*, *G. metallireducens*, and *S. oneidensis* during the response to various environmental stresses;
5. generate mutants in the various genes whose gene products are found to be responsible for a particular stress response, and to compare transcript, protein, and metabolite profiles in the mutant and isogenic wild-type strains;
6. produce small-molecule inhibitors that will interrupt the interaction between key proteins in the stress response, and to compare transcript, protein, and metabolite profiles in the treated and untreated strains; and
7. biopan the environment for stress response pathways.

A diagram explaining how these goals relate to each other and to the overall GTL project is shown in Figure 3.1. In general, samples (from bioreactors or the environment, subjected to a stress or grown in the absence of stress) will be obtained from the Applied Environmental Microbiology Core. These samples will be analyzed to elucidate the changes in transcript, protein, and metabolite levels in response to a particular stress. The data from these analytical techniques is sent to the Computational Core where it is analyzed to determine potential stress response networks. The information about potential stress response networks is used to design genetic mutants or chemical inhibitors to further elucidate the stress response. These chemically-inhibited wild-type strains and mutant strains are then subjected to the stress, and their response is compared to that of the wild-type strain. Further, mutations and/or chemical inhibitors may be needed depending on the nature of the network controlling the stress response. The Functional Genomics Core is highly dependent on the Applied Environmental Microbiology Core and the Computational Core for a variety of products. To see where the interfaces are see Figure 1.9. For details with AEM Core see Figure 2.1.

3.2 Background and Significance

When a cell is abruptly changed to a new environment, the cell suffers stress, particularly if the change in environment involves a change in pH, temperature, salinity, oxygen, metal concentration, irradiation, etc. The result of the change to a new environment is a change in the expression of a subset of genes (downregulation of housekeeping genes and upregulation of stress response genes), degradation of proteins (52), onset of sporulation, and any number of other responses. This response allows the cell to adapt to its new environment and survive.

Various stress response mechanisms have been studied in a number of bacteria. In general, the stress response systems in most bacteria appear to be generally well conserved (45). For example, *groESL* and *lon* (proteases) appear to be universal. In genomes of closely related bacterial species, even the order of the stress response genes is conserved (45). In addition, homologues of the gene encoding the universal stress protein (*uspA*) have been found in nearly all sequenced bacteria.

An important class of prokaryotic adaptive response systems consists of two signal transduction proteins: a sensor histidine kinase that “senses” extracellular stimuli, and a cognate response regulator that mediates the proper cellular response at the level of transcription regulation. Such two-component regulatory systems serve as a basic stimulus-response coupling mechanism that allows organisms to detect and rapidly respond to environmental

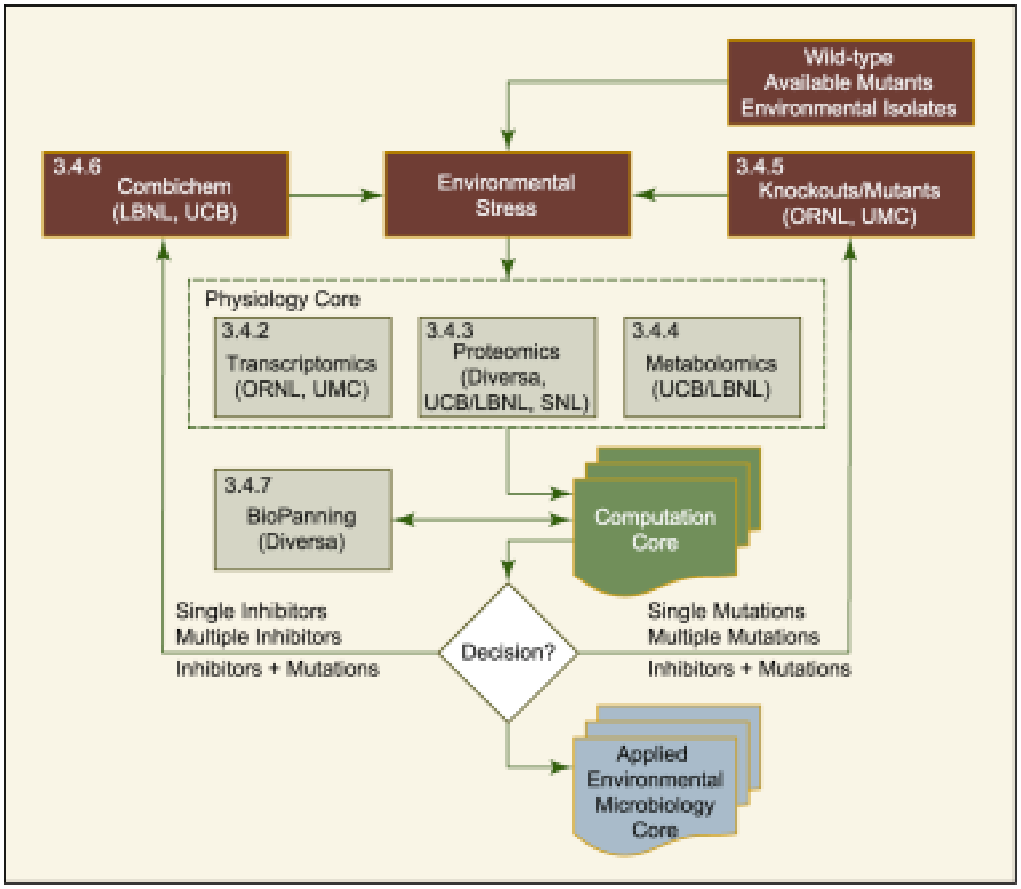


Figure 3.1. Outline of physiological monitoring.

changes. In bacteria, two-component regulatory systems govern the expression of target genes through controlled changes in protein phosphorylation [reviewed in (81)]. Signal reception alters the ability of a membrane-bound histidine kinase protein to transfer phosphate from ATP to a highly conserved histidine residue, typically located in the phosphoryl transfer-dimerization domain of the kinase protein. The transfer of phosphate from the histidine to an aspartate residue on the cognate response regulator changes the affinity of the latter protein for its target promoters. The phosphoryl transfer-dimerization domain of histidine kinase has been shown to specifically bind its cognate response regulator (94). The genomic sequencing of *D. vulgaris* Hildenborough, *S. oneidensis* MR-1, and *G. metallireducens* has revealed the presence of numerous putative two-component regulatory systems.

Genome-wide monitoring of the bacterial stress response is hardly a new area of research. Many research groups have used two-dimensional gels to analyze the response of various organisms to a number of stresses (57). While they were able to observe and later characterize the induction of several genes (by the appearance or disappearance of spots on 2-D protein gels), the analysis was tedious and difficult. The discovery of Magic Spot (ppGppp) has led to a number of metabolite profiling studies in which several phosphorylated metabolites were separated by 2-D thin-layer chromatography (TLC) and the most prominent spots analyzed (10). Again, it was relatively simple to see changes in the various spots on the TLC during the stress response, but it was more challenging to identify those spots. More recently, DNA arrays have been used to study the stress responses in *Escherichia coli* (63) and *Bacillus subtilis* (69), to name a few. The use of DNA arrays has significantly improved and eased analysis of the stress response, because one can readily associate the changes in the intensity of a particular spot on the array with the regulation of a particular gene. Improved methods for protein profiling and deducing protein-protein interactions will make proteomics a more useful technology for studying the stress response. Similar technologies are only now becoming available for profiling of a large fraction of the metabolite content of the cell.

However, to date, there have been no studies that use transcript, protein, and metabolite profiling together to obtain a more complete understanding of the stress response. We expect to gain new insight into the stress response by monitoring and comparing all three profiles simultaneously and at a significantly lower cost (in time and money) compared to the low-throughput approaches of the past.

The key genes involved in the stress response were discovered by mutating the genes and analyzing the response of the organism to the stress. Like much of the early work with 2-D gel electrophoresis and TLC, the construction of the mutants was tedious and the analysis was time-consuming. The completed genome sequences for our target organisms allow for the high-throughput construction of mutants in key stress response genes (and any other genes), which is feasible in a relatively short time. Further, the analysis of the response can be performed on a genome-wide scale, facilitating the discovery of genes that would not be discovered using traditional low-throughput approaches.

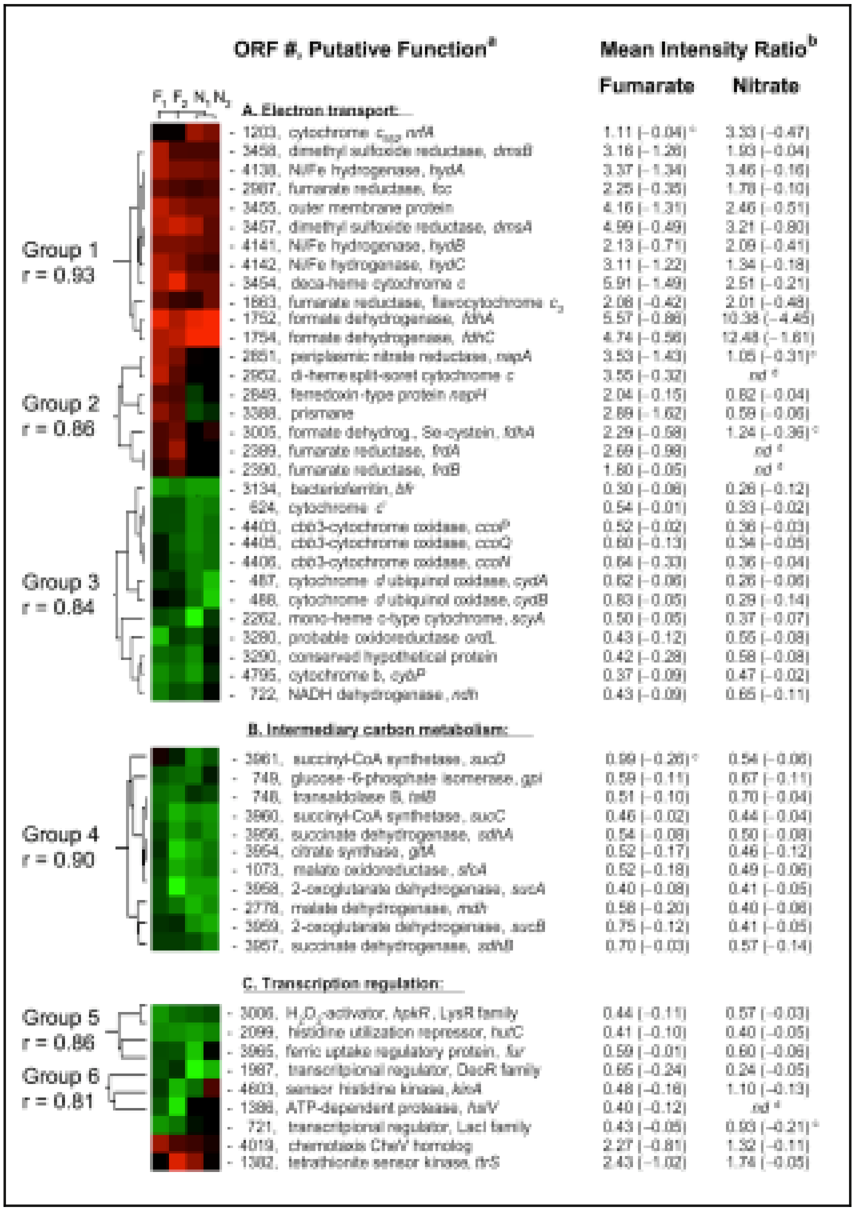
3.3 Preliminary Work

The researchers involved with the proposed work have experience with either the target organisms or the methods that will be used to investigate the stress responses. Thus, the preliminary work presented below may or may not pertain to the target organisms, but is presented as an example of the methods that will be used in the proposed work.

3.3.1 DNA Arrays of Shewanella oneidensis

Transcript profiling will be used to examine expression of genes that change during a stress. As such, we plan to use existing arrays (for *S. oneidensis*) or construct DNA arrays (for *D. vulgaris* and *G. metallireducens*) Under previous support from the DOE Microbial Genome (MGP) and NABIR Programs, we (J. Zhou's laboratory at ORNL) constructed DNA microarrays for *S. oneidensis* and initiated studies to identify genes and mechanisms involved in energy metabolism in MR-1. We have optimized and established protocols for microarray fabrication, total cellular RNA isolation, probe labeling, microarray hybridization, scanning, image processing, and data analysis. Using these procedures, we have demonstrated that microarray analysis can be used successfully to study gene function and regulatory mechanisms arising from sequence data and mutation studies. In our preliminary experiments we used partial-genome microarrays containing 691 genes to examine the putative role of the *S. oneidensis* Fnr-like *etrA* (electron transport regulator) gene in anaerobic gene regulation. Our results revealed altered mRNA levels for 69 genes putatively involved in energy metabolism, transcription regulation, substrate transport, and biosynthesis (6) (Figure 3.2). Among those, up to a 12-fold decrease in mRNA abundance was displayed by genes involved in anaerobic respiration (*dmsAB*, *hydABC*, *fdhAC*), while aerobic genes encoding cytochrome oxidases, NADH dehydrogenase, and TCA cycle enzymes were induced up to three-fold as a result of the *etrA* mutation. Notably, the mutation in the *etrA* locus affected the transcription of ten regulatory genes, including *fur* (ferric uptake regulator) and *hutC* (histidine utilization repressor). The sequence analysis identified putative FNR binding sites upstream of 27 genes displaying altered transcription levels in the etrA mutant. In addition, microarray analysis of a *fur* knockout strain of *S. oneidensis* resulted in a 3-fold reduction in *etrA* expression, as well as de-repression of genes involved in siderophore-mediated iron transport (83). These studies demonstrate the value of DNA expression microarrays for the analysis of genetic mutants.

We have also used microarrays to monitor differential gene expression in *S. oneidensis* under fumarate-, Fe(III)-, and nitrate-reducing conditions (5). In response to changes in redox and growth conditions, 121 genes of *S. oneidensis* displayed at least a 2-fold difference in transcript abundance (Figure 3.3). Genes induced during anaerobic respiration included those involved in cofactor biosynthesis and assembly (*moaACE*, *ccmHF*, *cysG*), substrate transport (*cysUP, cysTWA, dcuB*), and anaerobic energy metabolism (*dmsAB*, *psrC*, *pshA*, *hyaABC*, *hydABC*). A transcription of genes encoding a periplasmic nitrate reductase (*napDAGHB*), cytochrome c_{552}, and prismane was



ORF #, Putative Function[a]	Mean Intensity Ratio[b]	
	Fumarate	Nitrate
A. Electron transport:		
1203, cytochrome c_{552} *nrfA*	1.11 (±0.04)[c]	3.33 (±0.47)
3458, dimethyl sulfoxide reductase, *dmsB*	3.16 (±1.26)	1.93 (±0.04)
4138, NiFe hydrogenase, *hydA*	3.37 (±1.34)	3.46 (±0.16)
2987, fumarate reductase, *fcc*	2.25 (±0.35)	1.78 (±0.10)
3455, outer membrane protein	4.16 (±1.31)	2.46 (±0.51)
3457, dimethyl sulfoxide reductase, *dmsA*	4.99 (±0.49)	3.21 (±0.80)
4141, NiFe hydrogenase, *hydB*	2.13 (±0.71)	2.09 (±0.41)
4142, NiFe hydrogenase, *hydC*	3.11 (±1.22)	1.34 (±0.18)
3454, deca-heme cytochrome *c*	5.91 (±1.49)	2.51 (±0.21)
1883, fumarate reductase, flavocytochrome c_3	2.08 (±0.42)	2.01 (±0.48)
1752, formate dehydrogenase, *fdhA*	5.57 (±0.88)	10.38 (±4.45)
1754, formate dehydrogenase, *fdhC*	4.74 (±0.56)	12.48 (±1.61)
2851, periplasmic nitrate reductase, *napA*	3.53 (±1.43)	1.05 (±0.31)[c]
2852, di-heme split-soret cytochrome *c*	3.55 (±0.32)	*nd* [d]
2849, ferredoxin-type protein *napH*	2.04 (±0.15)	0.82 (±0.04)
3388, prismane	2.89 (±1.62)	0.59 (±0.06)
3005, formate dehydrog., Se-cystein, *fdhA*	2.29 (±0.58)	1.24 (±0.36)[c]
2389, fumarate reductase, *frdA*	2.69 (±0.98)	*nd* [d]
2390, fumarate reductase, *frdB*	1.80 (±0.05)	*nd* [d]
3134, bacterioferritin, *bfr*	0.30 (±0.06)	0.26 (±0.12)
624, cytochrome *c'*	0.54 (±0.01)	0.33 (±0.02)
4403, cbb3-cytochrome oxidase, *ccoP*	0.52 (±0.02)	0.36 (±0.03)
4405, cbb3-cytochrome oxidase, *ccoQ*	0.60 (±0.13)	0.34 (±0.05)
4406, cbb3-cytochrome oxidase, *ccoN*	0.64 (±0.33)	0.36 (±0.04)
487, cytochrome *d* ubiquinol oxidase, *cydA*	0.62 (±0.06)	0.26 (±0.06)
488, cytochrome *d* ubiquinol oxidase, *cydB*	0.83 (±0.05)	0.29 (±0.14)
2262, mono-heme c-type cytochrome, *scyA*	0.50 (±0.05)	0.37 (±0.07)
3280, probable oxidoreductase *ordL*	0.43 (±0.12)	0.55 (±0.08)
3290, conserved hypothetical protein	0.42 (±0.28)	0.58 (±0.08)
4795, cytochrome b, *cybP*	0.37 (±0.09)	0.47 (±0.02)
722, NADH dehydrogenase, *ndh*	0.43 (±0.09)	0.65 (±0.11)
B. Intermediary carbon metabolism:		
3961, succinyl-CoA synthetase, *sucD*	0.99 (±0.26)[c]	0.54 (±0.06)
749, glucose-6-phosphate isomerase, *gpi*	0.59 (±0.11)	0.67 (±0.11)
748, transaldolase B, *talB*	0.51 (±0.10)	0.70 (±0.04)
3960, succinyl-CoA synthetase, *sucC*	0.46 (±0.02)	0.44 (±0.04)
3956, succinate dehydrogenase, *sdhA*	0.54 (±0.08)	0.50 (±0.08)
3954, citrate synthase, *gltA*	0.52 (±0.17)	0.46 (±0.12)
1073, malate oxidoreductase, *sfcA*	0.52 (±0.18)	0.49 (±0.06)
3958, 2-oxoglutarate dehydrogenase, *sucA*	0.40 (±0.08)	0.41 (±0.05)
2778, malate dehydrogenase, *mdh*	0.58 (±0.20)	0.40 (±0.06)
3959, 2-oxoglutarate dehydrogenase, *sucB*	0.75 (±0.12)	0.41 (±0.05)
3957, succinate dehydrogenase, *sdhB*	0.70 (±0.03)	0.57 (±0.14)
C. Transcription regulation:		
3006, H_2O_2-activator, *hpkR*, LysR family	0.44 (±0.11)	0.57 (±0.03)
2099, histidine utilization repressor, *hutC*	0.41 (±0.10)	0.40 (±0.05)
3965, ferric uptake regulatory protein, *fur*	0.59 (±0.01)	0.60 (±0.06)
1987, transcriptional regulator, DeoR family	0.66 (±0.24)	0.24 (±0.05)
4603, sensor histidine kinase, *hksA*	0.48 (±0.16)	1.10 (±0.13)
1386, ATP-dependent protease, *hslV*	0.40 (±0.12)	*nd* [d]
721, transcriptional regulator, LacI family	0.43 (±0.05)	0.93 (±0.21)[c]
4019, chemotaxis CheV homolog	2.27 (±0.81)	1.32 (±0.11)
1382, tetrathionite sensor kinase, *ttrS*	2.43 (±1.02)	1.74 (±0.05)

Figure 3.2. Hierarchical clustering of the 69 *S. oneidensis* MR-1 genes exhibiting altered expression patterns in the *etrA* mutant. Each column indicates a separate biological replicate (F1 and F2 refer to fumarate-reducing conditions; N1 and N2 refer to nitrate-reducing conditions). Red and green colors indicate genes induced or repressed in the presence of *etrA*, respectively. The Pearson correlation coefficients (r) are displayed to the left of the nodes. [a]Sequence annotation was provided courtesy of TIGR (unpublished results). The ORF numbers provided here are for the purpose of tracking gene locations and annotation. The ORF numbers may change once the complete version of *S. oneidensis* MR-1 genome sequence annotation is released. [b]Relative expression is presented as the ratio of the dye intensity of the wild type to that of the *etrA* mutant averaged across two biological replicates. The standard deviation for each mean expression ratio is provided in parenthesis. [c]The absolute value of the mean ratio is not significantly different from 1 at $p = 0.05$ based on a one-tailed t-test. [d]Expression ratio cannot be determined due to the very low levels of hybridization intensities in both parental and ETRA1 strains.

elevated 8- to 56-fold in response to the presence of nitrate, while *cymA*, *ifcA*, and *frdA* were specifically induced 3- to 8-fold under fumarate-reducing conditions. The mRNA levels for two oxidoreductase-like genes of unknown function and several cell-envelope genes involved in multidrug resistance increased specifically in response to Fe(III) citrate reduction.

Our studies suggest that the proposed use of microarray-based methods will provide meaningful information about the induction and regulation of specific *S. oneidensis* genes involved in the reduction of different forms of metals. This approach will be especially useful for studying cellular responses under environmentally relevant conditions in the presence of multiple electron acceptors.

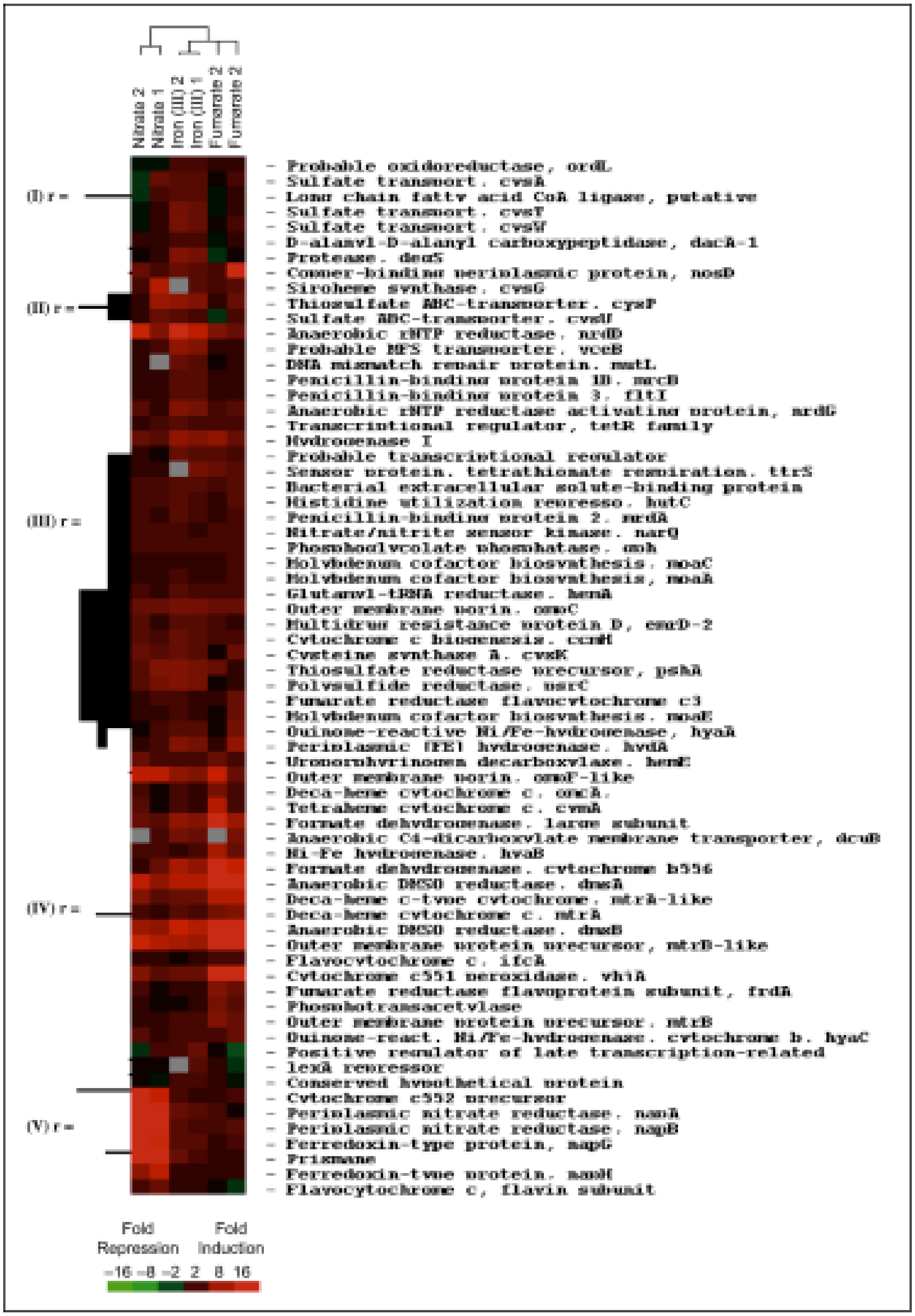


Figure 3.3. Hierarchical clustering of the 121 selected *S. oneidensis* MR-1 genes exhibiting altered expression patterns during the switch from aerobic to anaerobic growth. Each column indicates a separate biological replicate. (A) Clustering of genes displaying increased expression levels under anaerobic conditions. (B) Clustering of genes displaying decreased expression levels under anaerobic conditions. Red indicates up-regulation, green represents down-regulation, gray indicates missing data points. The Pearson correlation coefficients (r) are displayed to the left of the nodes. We selected seven gene clusters (I through VII), in which the pairwise correlations were greater than 0.90.

3.3.2 Proteomics

In addition to transcript profiling, we plan to monitor changes in the protein profile of our target organisms during the application of a stress. A key technology to be used to accomplish the proposed work is Diversa’s non-gel based proteomics method (32, 48, 72, 91, 96-98), which allows for the simultaneous identification and quantification (31-33) of individual proteins in complex mixtures, and overcomes the limitations inherent in traditional protocols (Figure 3.4). To achieve high-throughput qualitative proteome identification, we directly couple multidimensional □LC separation with tandem mass spectrometry to analyze the proteome in a fully automated manner. In detail, a denatured and reduced protein mixture is digested with a specific protease to produce desired peptide fragments. This complex peptide mixture is subsequently separated using a three-dimensional (3-D) microcapillary column containing reversed phase (RPC), strong cation exchange (SCX), and reversed phase (RPC), in that order. Without desalting, the peptide mixture is loaded directly onto a 3-D microcapillary column generated in-house at Diversa using a reliable and consistent procedure. A discrete fraction of the absorbed peptides is displaced from the first RPC to the SCX section using a reverse phase gradient (X_n–X_{n+1}%). This fraction of peptides is retained onto the SCX section and then subfractionated from the SCX column onto the RPC column using a step gradient of salt, where part of the peptides are eluted and retained on the last RPC section while contaminating salts and buffers are washed through. The subfractionated peptides are then separated on the RPC column using the same reverse phase gradient (X_n–X_{n+1}%). The masses and sequences of separated and eluted peptides are detected directly by a tandem mass spectrometer. This process is repeated using increasing salt concentration to displace additional subfractions from the SCX column, following each step by a reverse phase gradient. Upon completion of the whole sequence of salt steps, the process is repeated, employing a higher reverse phase gradient (X_{n+1}–X_{n+2}%, X_{n+2}> X_{n+1}, n=0, 1, 2, 3…, X_1=0). Each of the cycles is applied in an iterative manner, with the total number of cycles depending on the complexity of the peptides. In general the processing of a complex protein mixture involves 3–6 acetonitrile cycles followed by 3–6 salt gradient steps. The MS/MS data from all of the fractions are analyzed by database searching.

3-D LC MS is a fully automated technique using LC in combination with mass spectrometry and database search for highly complex mixtures. It is universal, and it identifies proteins with extreme pI, any MW, and a wide variety of protein classes. In contrast to conventional 2-D gel methods it can access hydrophobic proteins. It has high sensitivity, peak capacity, and gives dynamic range greater than 10,000 to 1.

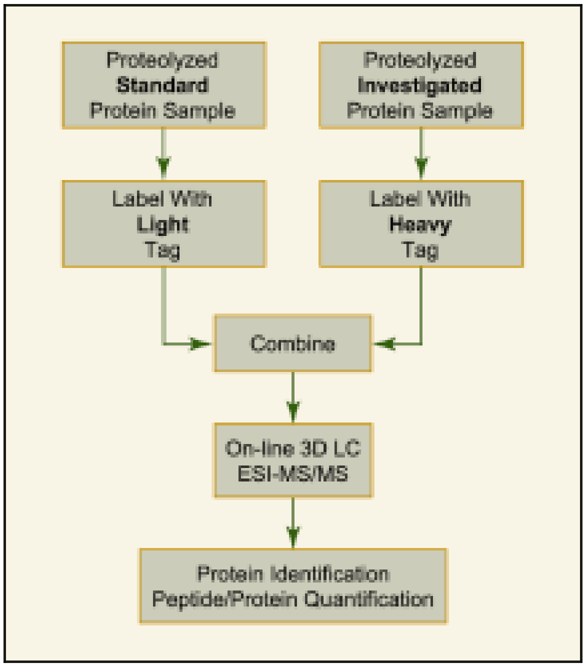
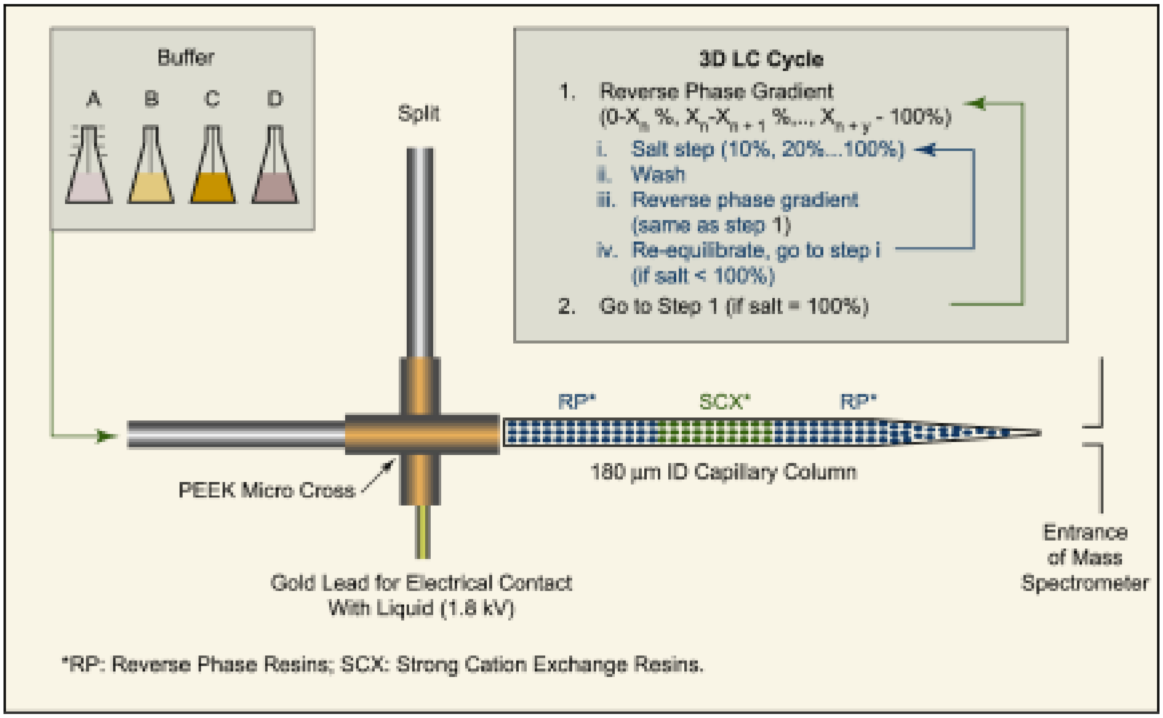

Proteolyzed Standard Protein Sample
Proteolyzed Investigated Protein Sample
Label With Light Tag
Label With Heavy Tag
Combine
On-line 3D LC ESI-MS/MS
Protein Identification Peptide/Protein Quantification

Figure 3.4. Schematic representation of high-throughput quantitative proteomics process.

To achieve comparative profiling proteomics, we combine 3D LC with a novel differential tagging method developed at Diversa. Specifically, we compare two or more samples of proteins, the first of which is designated as the standard sample, and all others designated as test samples. First, the proteins in the standard and test samples are subjected to a sequence of proteolytic digestion in separate tubes. The digested peptides are then differentially modified using a chemical method, such as esterification of C-termini of the peptides and carboxylic acid groups in the side chains, for example, D0/D5 ethoxylation. Peptides derived from the standard and the test samples are labeled with chemical residues having similar overall properties but different masses. The samples are then combined, separated by 3-D chromatography, and analyzed by mass spectrometry methods. Finally, mass spectrometry data are processed by special software, which allows for identification and quantification of proteins. A schematic illustration of this procedure is shown in Figure 3.5.

3.3.2.1 High-throughput Proteomics. The proposed work involves elucidating the protein profiles for three different organisms and several different stress conditions. Because of the large number of samples that must be analyzed, we plan to develop and use high-throughput proteomic techniques. Researchers at SNL have developed capillary- and microchip-based separations for biomolecules, including peptides and proteins (78, 84, 95). A multicapillary separation system (Figure 3.6), analogous to 96-capillary devices for separation of DNA, can potentially be used for high throughput and fast analysis of proteins. Each capillary in the array can operate at a different solvent or pH, or can use a different chromatographic phase, leading to a different separation of the same protein mixture. Data-analysis algorithms can then be used to extract information from the multiplexed system. Alternatively, multiple identical separation capillaries can be used to study the effect of different stressors on the same microbe, and results on the differential expression can be used to elucidate metabolic pathways. The separation options implemented in capillaries include electrophoresis, gel electrophoresis, isoelectric focusing, electrochromatography, and HPLC.


Buffer
A
B
C
D
Split
3D LC Cycle
1. Reverse Phase Gradient
(0-X_n %, X_n-X_{n+1} %,..., X_{n+y} - 100%)
i. Salt step (10%, 20%...100%)
ii. Wash
iii. Reverse phase gradient
(same as step 1)
iv. Re-equilibrate, go to step i
(if salt < 100%)
2. Go to Step 1 (if salt = 100%)
RP*
SCX*
RP*
180 µm ID Capillary Column
PEEK Micro Cross
Gold Lead for Electrical Contact
With Liquid (1.8 kV)
Entrance
of Mass
Spectrometer

*RP: Reverse Phase Resins; SCX: Strong Cation Exchange Resins.

Figure 3.5. Mixed-bed 3-D LC separation.

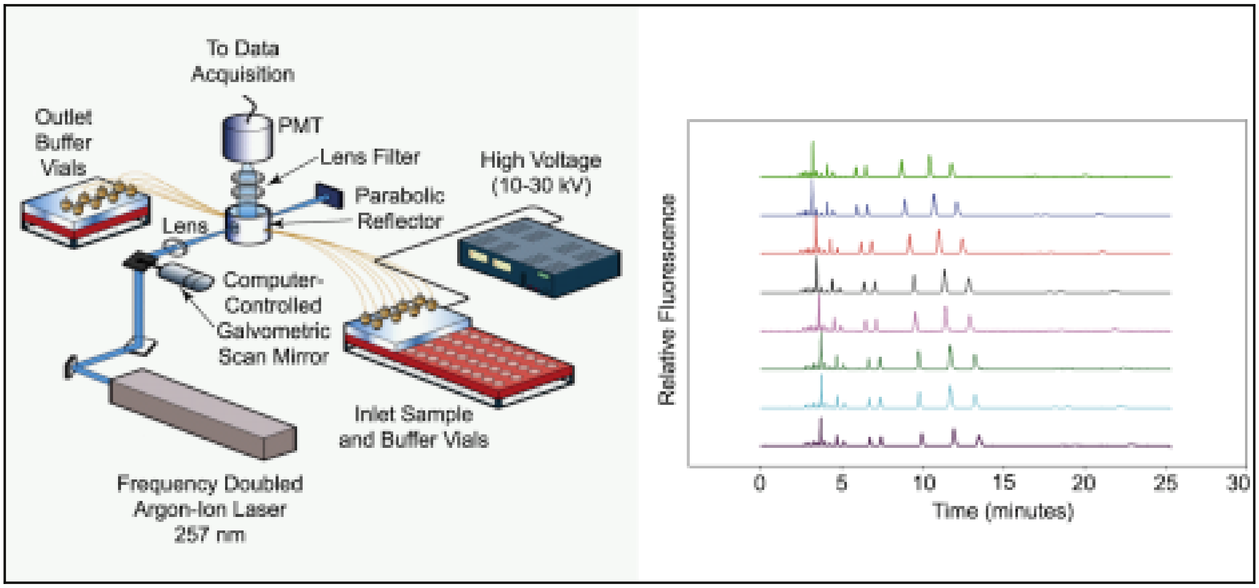


Figure 3.6. Capillary- and microchip-based system for protein separations and analysis. (Left) Device. (Right) Typical separation.

3.3.3 Metabolomics

As mentioned in Subsection 3.2, several metabolites change dramatically during stress responses. In a few cases, these metabolites regulate gene expression involved in the stress response. As such, we propose to investigate the metabolite profile of the cell during the response to a particular stress. Not only is it critical to identify the metabolites that change during the stress response, it is also critical to asses their quantity (pool sizes) and their turnover (flux). The Keasling laboratory has developed a number of methods to assess fluxes through the various metabolic pathways in the cell (65, 67, 68). They have formulated an analytical framework from the annotated genome sequence (which reveals the metabolic pathways present in the cell), the composition of the cell, and any measured consumptions, excretions, or intracellular fluxes. At steady state, one can solve for the intracellular fluxes using any number of mathematical methods. Since the model predicts how the fluxes through the metabolic pathways must be distributed in order to achieve a specific growth rate and product, one can compare the fluxes through the metabolic pathways under various growth conditions and predict which metabolic pathways will be regulated. In addition, one can predict how alterations in metabolic pathways will affect growth of the cell.

The sensitivity of the model to changes in fatty acid composition and amino acid composition was tested using data collected under batch and continuous growth conditions. Correlations describing how the total amounts of the cellular macromolecules change with growth rate were developed from experimental data and were used to calculate the drain of precursors from the metabolic network for the synthesis of macromolecules, coenzymes, and prosthetic groups at a specific growth rate. The precursor composition of two key macromolecules, the amino acid composition of proteins, and the fatty acid composition of the cell membranes were measured in cells growing exponentially in batch cultures on glucose, succinate, glycerol, pyruvate, and acetate, and in cells growing under continuous culture conditions on glucose (Figure 3.7) at dilution rates equivalent to the growth rates of the batch cultures. The amino acid content of proteins did not change significantly with growth condition (carbon source) or dilution rate. However, the fatty composition of the membranes did change significantly, both with carbon source and dilution rate. The underdetermined set of equations was solved using the Simplex algorithm, employing realistic objective functions and constraints; the drain of precursors, coenzymes, and prosthetic groups, and the energy requirements for the synthesis of macromolecules, served as the primary set of constraints. The model results accurately predicted a known regulation for growth on various carbon sources. Similar techniques, used in conjunction with isotopomer labeling of intracellular metabolites, will be used to calculate intracellular fluxes of

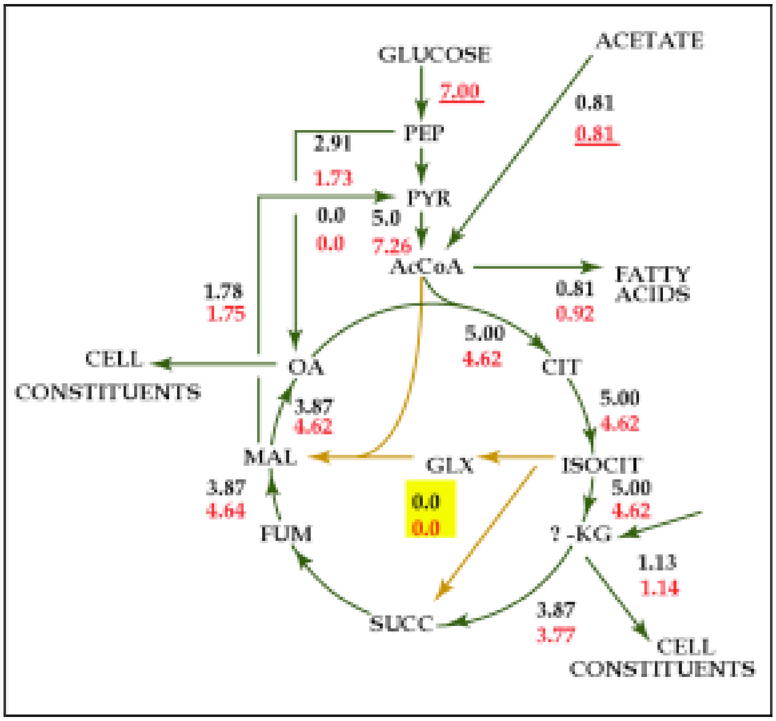


Figure 3.7. Comparison of measured and computed fluxes in *E. coli* during growth on glucose. The numbers in black (top) are experimentally measured fluxes, and the numbers in red (below) are calculated fluxes.

metabolites, particularly those such as ppGppp and others that are known to be key components of the stress response.

3.3.4 Combinatorial Synthesis of Small-Molecule Inhibitors: Disruption of Sulfate-Reduction Pathways in Bacteria

The technique of combinatorial small-molecule library synthesis has enabled the rapid generation of large compound libraries for screening against biological targets, and is now widely applied by pharmaceutical companies to drug-discovery efforts. We plan to adapt combinatorial library strategies to the rapid identification of chemical inhibitors of molecular interactions involved in environmental stress response pathways. The proposed use of small molecules to inhibit key components in the stress response pathways has many potential advantages. In addition to being able to duplicate many of the results currently being developed from gene-knockout studies, small-molecule inhibitors have the potential to provide more complete information about the various relationships of the metabolic systems, their mutual dependencies, their regulation, and their networks. This lack of dependence upon conventional gene-knockout methods is a unique aspect of this proposal, and could lead to major breakthroughs in the study of microbial metabolic networks.

Some of the technical advantages of the small-molecule approach are derived from the high level of control that one can exercise over the means of disrupting a gene product. For example, small-molecule inhibitors can be dosed at various levels, rather than simply being used as an on/off switch. This is in contrast to gene knockouts that can be lethal to the organism in many cases. At the same time, many systems can be modulated in a controlled manner—either in parallel or in series. The molecular inhibitors can act on a diffusion-limited time scale to shut down a protein's activity before other systems have time to adjust and mask the effect. This element of temporal control is lacking with most genetic methods. In addition, chemical inhibition is reversible. These properties make it possible to perturb these essential systems transiently and at various levels to study their reverberation into other metabolic pathways.

Using the combinatorial-chemistry approach of a small-molecule library, Professor Bertozzi's group has recently identified some potential small-molecule inhibitors targeted against the tuberculosis APS reductase. In this study we functionally defined the sulfate assimilation pathway of several soil-resident mycobacteria, cloned two genes (*cysH* and *cysN/cysC*) involved in the manipulation of APS (encoding APS kinase and APS reductase, respectively), and have characterized their specificity by genetic complementation in *E. coli* strains (JM81A) defective in known places in their sulfate assimilation pathway (Figure 3.8) (77). We used this complementation-based screening approach to search for inhibitors of mycobacterial APS kinase and APS reductase. Strains bearing complementation plasmids were grown in M9 minimal media in 384 well plates. Using the high-throughput screening facility at the Institute of Chemistry and Chemical Biology at Harvard University, 18,000 compounds were added to each of the two complemented strains and to the control for a total of 54,000 experiments. Compounds were transferred using robotic pin-transfer into each of the 384 well plates for a final concentration of 12–25 mg L^{-1}. Cells were then grown at 37ºC for two days before measuring their absorbance at 650 nm on a 384-well plate reader. Absorbance values for the experimental strains were converted into percentage inhibition relative to the reference strain. Fifty compounds that gave a 40% or greater inhibition of growth on one or another experimental strain were found, but not on the control strain. These 50 compounds were cherry-picked, and the inhibition assay was repeated on a larger scale to confirm the observed phenotype. Figure 3.9 shows four of the most potent compounds detected so far. Further work is planned to test each of these compounds at different concentrations to ensure a dose-dependent response, and then to establish in vitro assays to confirm their ability to act directly on the target enzyme. We will use a similar approach to identify small molecule inhibitors of the various stress response regulatory pathways.

3.4 Research Design and Methods

The goal of this section is to develop the experimental methods to elucidate the regulatory networks in the stress response of *D. vulgaris, S. oneidensis*, and *G. metallireducens.* We will develop methods (transcriptomics, proteomics, metabolomics) to analyze the physiological response of these organisms to various stresses. To elucidate the genes and proteins responsible for sensing a particular stress and initiating a stress response, we will use (and develop, in some cases) mutagenesis techniques to mutate genes involved in the stress response and small molecule inhibitors to inhibit two-component regulatory interactions that initiate the response to stress.

3.4.1 *Growth Conditions and Stress Applications*

The target organisms will be grown by the Applied Environmental Microbiology Core investigators (described in Section 2.4.4). In addition, they will apply the appropriate stresses to the organism. Samples will be collected according to the guidelines prescribed by each group in the Functional Genomics Core and shipped to the various locations for subsequent analysis.

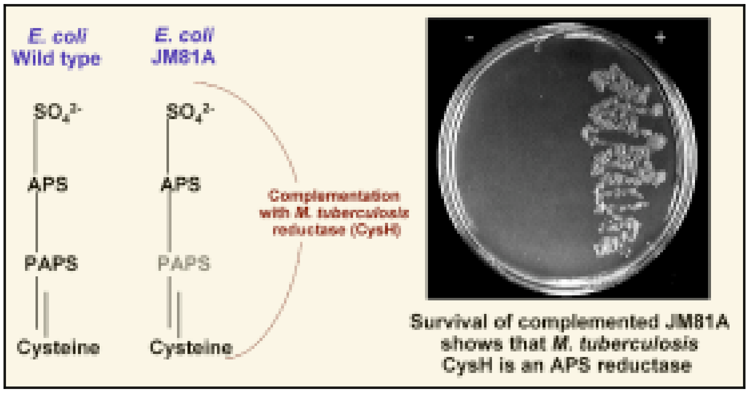

E. coli Wild type
E. coli JM81A
SO4 2-
SO4 2-
APS
APS
PAPS
PAPS
Cysteine
Cysteine
Complementation with M. tuberculosis reductase (CysH)
-
+
Survival of complemented JM81A shows that M. tuberculosis CysH is an APS reductase

Figure 3.8. Complementation of *E. coli* JM81A by *M. tuberculosis* CysH shows that this gene is an APS reductase. Cells grown on minimal media as described in the methods section. (-) JM81A no plasmid. (+) JM81A containing a constitutive expression plasmid bearing the *M. tuberculosis* CysH gene.

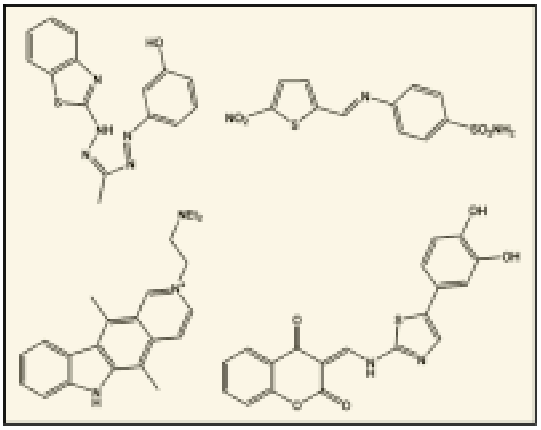


Figure 3.9. Compounds identified as inhibitors of mycobacterial APS kinase and APS reductase using complementation-based screening.

3.4.2 Transcript Profiling

To delineate the genes and identify the regulatory systems implicated in the cellular response to different types of stress, we will compare mRNA patterns expressed under normal physiological (nonstress) conditions to those preferentially exhibited under stress conditions. The stress factors to be tested will include oxygen tension (for *D. vulgaris* and *G. metallireducens* only), pH, phosphate concentrations, salinity, nutrient availability, and metal toxicity (Cd, Hg, Ni, and Cr). Genes in *D. vulgaris*, *S. oneidensis*, and *G. metallireducens* that are differentially expressed in response to these environmental factors will be displayed on a genomic scale using high-density, whole-genome DNA microarrays. The tasks in this part of the proposal are (1) to construct whole-genome microarrays for *D. vulgaris*, *S. oneidensis*, and *G. metallireducens*; (2) to define the global transcriptional response of *D. vulgaris*, *S. oneidensis*, and *G. metallireducens* to various environmental stress conditions using whole-genome DNA microarrays; and (3) to analyze the transcriptional response of mutants and small molecule-inhibited *D. vulgaris*, *S. oneidensis*, and *G. metallireducens*.

3.4.2.1. Whole-Genome Microarrays for *D. vulgaris*, *S. oneidensis*, and *G. metallireducens*

***Shewanella* Microarray.** To investigate alterations in gene-expression profiles in response to stress, we will use whole-genome DNA microarrays for *S. oneidensis*, which are close to being completed at ORNL (Dr. J. Zhou's laboratory) under previous support from the Department of Energy's MGP and NABIR programs. The array elements consist of PCR-amplified DNA fragments corresponding to 4,498 predicted open reading frames (ORFs) and 285 oligonucleotides (50 mers), representing approximately 97% of the total protein-coding capacity of the *S. oneidensis* MR-1 genome. The oligonucleotides were designed for ORFs that could not be amplified by PCR using our reaction conditions. PCR primers for the amplification of putative genes from *S. oneidensis* MR-1 were designed using the computer program Primer 3 (Whitehead Institute, Cambridge, MA) with genome sequence information provided as a courtesy of TIGR (TIGR, unpublished results). PCR amplification was considered to be successful if reactions contained a single product of the expected size as determined by agarose gel electrophoresis and ethidium bromide staining. For array deposition, PCR products and presynthesized oligonucleotides were diluted in 50% DMSO (Sigma Chemical Co., St. Louis, MO) to a final concentration of 100–300 ng/□l. For control and normalization purposes, the following samples were also included in the array: (i) a set of six serial dilutions of *S. oneidensis* MR-1 genomic DNA as positive controls; (ii) 10 unique *Arabidopsis thaliana* genes from the Stratagene SpotReport Kit (Stratagene, La Jolla, CA); and (iii) blank control spots containing 50% DMSO. These controls are present on each of the array's 16 subgrids, while no two controls are adjacent to each other. The DNA elements (PCR products, oligonucleotide probes, and controls) will be printed onto, and chemically linked to, amino-modified glass slides (Telechem International, Sunnyvale, CA) with ChipMaker III pins (Telechem International) using a PixSys 5500 robotic printer (Cartesian Technologies Inc., Irvine, CA). Slides will be postprocessed according to the manufacturer's instructions (Telechem International).

***Desulfovibrio* Microarray.** *Desulfovibrio vulgaris* strain Hildenborough is a □*-proteobacterium* with a genome size of 3.6 Mb and a G+C genome content of 65%. The genome sequencing of this bacterium has recently been completed, and the gaps have been closed (see the following Web site for sequence data and annotation: http://www.tigr.org/tigr-scripts/ufmg/ReleaseDate.pl). We will use the finalized sequence annotation of *D. vulgaris* provided by TIGR to design primers for all the genome ORFs using the same primer program that was used for *S. oneidensis* MR-1. The fabrication of the whole-genome DNA microarray for *D. vulgaris* will be carried out as described above. It is important to note that a proposal aimed at constructing *D. vulgaris* Hildenborough microarrays has been recently submitted to the NABIR program (Drs. Matthew Fields and Judith Wall). If funded, whole-genome array fabrication for this organism will be carried out under NABIR support.

***Geobacter* Microarray.** Under the support of DOE, the whole-genome sequencing for *G. metallireducens* is under way at the Joint Genome Institute (Walnut Creek, CA). It is expected that the high-draft version of the *G. metallireducens* genome will be released within the next two months (by June 2002). A genome annotation that will be used to construct a whole-genome microarray of *G. metallireducens* will be prepared at ORNL by the Computational Biology Group under the direction of Frank Larimer. The array fabrication will be carried out as described above.

3.4.2.2. Global Transcriptional Response of *D. vulgaris* Hildenborough, *S. oneidensis* MR-1, and *G. metallireducens* to Various Environmental Stresses Using Whole-Genome DNA Microarrays. Total cellular RNA will be isolated immediately using a standard protocol (5) from the cultures subjected to stress. cDNA probes will be synthesized from 10 □g of total RNA using random nonamer primers and the recombinant M-MLV Super-Script II RNase H$^-$ reverse transcriptase (Invitrogen) in the presence of amino-allyl modified dUTP (Sigma). The amino allyl-labeled first-strand cDNA will be purified using a PCR purification kit (Qiagen) and will be subsequently conjugated with monoreactive cyanine dye, Cy3 or Cy5 (Amersham). For differential display of gene expression, the cDNA prepared from control cultures will be labeled with Cy3, whereas the cDNA from test cultures exposed to different stress factors will be labeled with Cy5. Hybridization and scanning will be carried out using protocols established in our previous studies (5, 83). Microarray slides will then be scanned using the ScanArray 5000 system (Packard Biochip) at 10 □m per pixel resolution.

Image analysis for determining fluorescence and background intensity and for flagging spots with poor quality will be performed using the microarray quantitation software ImaGene 4.2.1 (BioDiscovery). The following criteria for assessing spot integrity and quality will be used by all investigators, regardless of the image analysis software. First, to identify spots with uneven intensities, spots with more than 55% of pixels greater than one standard deviation (SD) above local background will be flagged as poor-quality spots. Then, to identify spots with signals not significantly above background, spots with median signals less than two times the standard deviation above background in both channels (11, 35) will be flagged as poor-quality spots and not included in data analysis.

To compare gene-expression profiles under various growth conditions, the data will be normalized and analyzed with the microarray data-analysis software package GeneSpring (Silicon Genetics). Hierarchical cluster analysis using both parametric and nonparametric approaches will be performed to display clusters of related gene-expression profiles. Other multivariate analysis, such as self-organizing maps (82) and principal-component analysis, will be used to identify groups of genes with common expression patterns. These analyses will also provide valuable information for assigning potential gene function to hypothetical ORFs.

The second component of the physiological monitoring of the stress response entails describing the response of global cellular proteins to a particular stress and elucidating the interactions of some of the key signal-transduction proteins responsible for sensing and invoking the response.

3.4.3 Protein Profiling

3.4.3.1. Global protein profiles of *D. vulgaris, S. oneidensis, and G. metallireducens*. To qualitatively and quantitatively characterize the protein profiles of *D. vulgaris, S. odeidensis,* and *G. metallireducens* during the various stress responses, we will use the protocols described in Subsection 3.3.2. Specifically, we will isolate soluble and membrane-associate proteins, denature and digest the fractions, and then esterify them using either D0 or D5 ethanol. Samples from two conditions (i.e. stress and unstressed) will be combined, separated, and analyzed by the (3-D) □LC MS/MS system, as described above. The acquired MS/MS data will be processed by special software that allows for identification and quantification of proteins.

This analysis will be performed on *D. vulgaris*, *G. metallireducens*, and *S. odeinesis* subjected to a number of stress conditions. In addition, we propose to perform this analysis on mutant strains of each of these organisms and on wild-type strains subjected to various small molecule inhibitors.

3.4.3.2. Identification of Binding Partners for Proteins Involved in the Transcriptional Regulation of the Cellular Response to Stress Using Phage Display. As mentioned in Subsection 3.2, two-component regulatory systems sense and initiate the response to a particular stress. We propose to determine which proteins are involved in that sensing process and to characterize their interactions (Table 3.1). To identify the binding partners for proteins implicated in the transcriptional control of stress responses, with the emphasis on defining the cognate pairs of histidine kinase and response regulator proteins of two-component systems, we will use a high-throughput phage display–based method. This research will build upon and complement phage display work proposed in the Microbial Cell Project for *S. oneidensis* MR-1. The proposed study will focus on (i) investigating the protein members of the signal transduction cascades and regulatory circuits underlying cellular responses to environmental stress and (ii) identifying the cognate response regulators for the sensor histidine kinases. Specifically, we will (1) clone the entire set of MR-1 open reading frames into a universal vector and then select target proteins for recombination with His-tag or GST-tag pHOST vectors; (2) screen the *S. oneidensis* phage display library in order

Table 3.1. List of putative genes from S. oneidensis MR-1 implicated in detoxification and general stress response (arranged in alphabetical order).

ORF #	Putative function (TIGR annotation, July 2001)
DETOXIFICATION:	
ORF01272	AhpC (ahpC)
ORF01274	Alkyl hydroperoxide reductase subunit (trxB)
ORF04793	Antioxidant, AhpC/Tsa family
ORF04646	Arsenate reductase (arsC)
ORF00307	Catalase
ORF01134	Catalase
ORF01551	Catalase/peroxidase HPI
ORF02822	Catalase/peroxidase HPI
ORF02740	Cation efflux family
ORF04634	Manganese superoxide dismutase Mn-SOD (sodB)
ORF01438	Multidrug efflux pump channel protein
ORF01251	Organic hydroperoxide resistance protein
ORF03740	Organic solvent tolerance protein precursor (imp)
ORF02393	Thiophene and furan oxidation (thdF)
ORF04078	Thioredoxin peroxidase (tagD)
ORF00836	Transcriptional regulator, oxyR-like, LysR family
STRESS RESPONSE:	
ORF04708	Carbon starvation protein A, putative
ORF04709	Carbon starvation protein A, putative
ORF01038	Probable DNA-binding stress protein
ORF03690	Universal stress protein family
ORF03691	Universal stress protein family
ORF05258	Universal stress protein family

to enrich for specific binding partners; and (3) validate protein-protein interactions using the yeast two-hybrid system. Once we have determined that these interactions can be found using this system and have identified some of the two-component stress response regulators for *Shewanella*, we will perform similar studies with *Desulfovibrio* and *Geobacter.*

Systematic cloning of open reading frames into Univector plasmids and expression of target proteins by recombining relevant univector plasmids with His-tag and GST-tag pHOST vectors. Under previous support from the DOE Microbial Cell Project (J. Zhou), the entire complement of *S. oneidensis* MR-1 genes will be amplified by PCR and inserted into a pUNI vector by directional cloning. The structure and orientation of the inserts will be confirmed by PCR. Because the genes inserted into the pUNI vector are derived from PCR amplification, it is possible that the cloned ORFs could contain mutations. To mitigate the possibility of errors, a high-fidelity DNA polymerase will be used for PCR amplification. In addition, the initial 100 ORF insert clones will be sequenced to provide an indication of the number of mutations present in the ORFs. If, as expected, the number of genes mutated is less than 1%, the clone set will be completed and used for conversion to phage display vectors. In this case, each of the pUNI plasmids in the clone set will be fused with the Fos-*loxP* plasmid to create a phage display library consisting of each open reading frame fused to Fos for Jun-Fos phage display. If the number of mutated genes is greater than 1%, an alternative polymerase will be used with more DNA template and fewer cycles to reduce mutations. We are currently in the process of constructing this univector plasmid library for *S. oneidensis* MR-1.

One of the advantages of the univector system is that a number of different pHOST vectors can be fused to the pUNI plasmid containing the gene of interest. The pHOST plasmid contains the appropriate promoter sequences or tag sequences (e.g., glutathione-S-transferase (GST) or polyhistidine) for creating fusion proteins. Candidate proteins to be used as the targets for phage display will include regulatory and sensor proteins identified by genomic sequence annotation and those gene products implicated in stress responses as determined by microarray analysis and mutagenesis. In addition to candidates identified using microarray analysis and mutagenesis, we will select putative histidine kinases, response regulators, and GGDEF domain-containing proteins (Table 3.2). These target proteins will be expressed by recombining the relevant univector plasmids with His-tag and GST-tag pHOST vectors. Cre-*loxP* mediated site-specific recombination fuses the pUNI and pHOST plasmids at the *loxP* site (50), thus permitting the generation of protein fusions to poly-His or GST tags. The recombinant target protein will then be purified using the appropriate affinity matrix.

Pan the systematic *S. oneidensis* phage display library to enrich for specific binding partners. The purified target protein generated (as described above) will be immobilized in microtiter wells or linked to oxirane acrylic beads (38, 60). The systematic *S. oneidensis* phage library constructed under the support of the DOE MCP project will be used for panning to enrich proteins that specifically bind to the target fusion proteins. Clones that are identified by panning and sequencing will be tested for specific binding using phage ELISA. For these experiments, the target protein will be immobilized in a microtiter well, and the phage clone encoding the putative binding partner will be allowed to bind. After extensive washing, bound phage will be detected with an anti-M13 phage antibody conjugated to horseradish peroxidase (HRP) (38). As a control for specific binding, several other proteins will also be immobilized and tested for binding by the phage. A further test for specific binding will be to demonstrate that the addition of soluble target protein to the phage-binding reaction inhibits binding to the immobilized target. This phage display approach will allow us to identify any potential cross-reactivities between sensor kinases and response regulators of different two-component systems. Sequence similarities probably lead to structural similarities that are responsible for cross-talk observed in vitro between members of different two-component regulatory systems (25, 39, 59).

Validate protein-protein interactions using the yeast two-hybrid system. The protein interactions identified by phage display (as described above) will be validated using the yeast two-hybrid system. pHOST vectors are available to convert univector clones into "bait" and "prey" plasmids for the two-hybrid assay (50, 75) (Figure 3.10). Here, we will adapt and use a high-throughput procedure, which uses the popular Gal4 system developed for

Saccharomyces cerevisiae (85). For this, the univector genome clones will be fused to the Gal4 activator domain ("prey" constructs), while the ORFs of interest will be converted to "bait" molecules carrying the Gal4 DNA-

Table 3.2. List of putative S. oneidensis genes involved in two-component signal transduction (arranged by ORF number).

ORF #	Putative function (TIGR annotation, July 2001)
ORF00112	Two-component response regulator PhoP
ORF00113	Two-component sensor PhoQ
ORF00399	GGDEF family protein
ORF00464	GGDEF family protein
ORF00568	Phosphate regulon sensor protein PhoR (*phoR*)
ORF00569	DNA-binding response regulator PhoB (*phoB*)
ORF00770	Response regulator
ORF00837	Sensor histidine kinase FexB(*fexB*)
ORF00950	Sensor protein TorS, putative
ORF01235	Urease domain protein
ORF01246	Two-component sensor, putative
ORF01293	PAS domain S-box ?
ORF01312	Regulatory components protein? of sensory transduction system
ORF01393	Response regulator
ORF01394	Sensory transduction histidine kinase
ORF01468	Sensory box/GGDEF family protein
ORF01667	Sensory box/GGDEF family protein
ORF01673	Two-component system, regulatory protein
ORF01674	Two-component system, sensor protein
ORF01729	Sensory box sensor histidine kinase/response regulator
ORF01738	Response regulator
ORF01739	GGDEF family protein
ORF01767	Response regulator (*cheY-3*)
ORF01768	Sensory box sensor histidine kinase
ORF02327	Two-component sensor KdpD
ORF02328	Two-component response regulator KdpE
ORF02443	Sensor histidine kinase
ORF02525	Two-component system, sensor protein
ORF02526	Two-component system, regulatory protein (TrcR)
ORF02539	Probable two-component sensor
ORF02540	Two-component system, regulatory protein (TcrA)
ORF02544	Osmolarity sensor protein EnvZ
ORF02545	Transcriptional regulatory protein Ompr
ORF02556	Two-component system, regulatory protein (TcrA)
ORF02608	Probable two-component response regulator
ORF02723	Probable two-component sensor
ORF02724	Two-component system, regulatory protein (TrcR)
ORF02737	Sensor protein cpxa (cpxA)
ORF02774	Virulence sensor protein BvgS precursor, other roles
ORF02820	GGDEF family protein
ORF02920	Sensory box/GGDEF family protein
ORF03103	Probable two-component sensor
ORF03104	Two-component system, regulatory protein (DctD-2)
ORF03303	Response regulator
ORF03304	Sensory transduction histidine kinas
ORF03307	Regulatory component of sensory transduction system

ORF03327	GGDEF family protein
ORF03425	Sensory box sensor histidine kinase
ORF03466	GGDEF family protein

Table 3.2. List of putative S. oneidensis genes involved in two-component signal transduction (arranged by ORF number) (Continued).

ORF #	**Putative function (TIGR annotation, July 2001)**
ORF03634	Putative sulfur deprivation response regulator
ORF03682	Two-component response regulator (*flrC*)
ORF03790	Probable two-component sensor
ORF03947	Sensor histidine kinase/response regulator
ORF04203	Two-component sensor kinase
ORF04277	Sensory box sensor histidine kinase
ORF04311	Styrene sensor kinase (*dctB*)
ORF04625	PAS domain S-box protein
ORF04710	Response regulator
ORF04811	Probable two-component sensor
ORF05031	Response regulator (*cheY-4*)
ORF05033	Sensory box sensor histidine kinase (*dctB*)
ORF05034	Sensory box sensor histidine kinase/response regulator
ORF05035	Sensory box sensor histidine kinase (*dctB*)
ORF05037	Response regulator, putative
ORF05038	Response regulator receiver domain protein (CpxR)
ORF05040	Response regulator
ORF05075	Sensory box/GGDEF family protein, putative
ORF05444	GGDEF family protein
ORF05449	Sensory protein BvrS, putative
ORF05474	GGDEF family protein
ORF05511	Sensor histide kinase
ORF05533	Probable two-component response regulator
ORF05542	Response regulator receiver domain protein
ORF05557	Probable sensor protein YgiY
ORF05620	Sensory box/GGDEF family protein

binding domain and HIS3 reporter gene. The resulting “prey” constructs will be transformed into a strain of opposite mating type of a Gal4 DNA-binding domain vector. To screen for protein interactions, we will mate the transformant containing one of the DNA-binding domain hybrids (“bait”) to all of the transformants of the array (“preys”), selecting diploids using markers carried on the two-hybrid plasmid. The diploids will then be transferred to selective plates deficient in histidine, and colonies positive for the two-hybrid reporter HIS3 gene will be identified. For each of the screens, we will obtain on the order of up to 30 positives, since only around 20% of these positives are reproduced in a second screen (85). Although the exact causes of this variability are not known, they appear to include an infrequent and protein-specific rearrangement of the DNA-binding domain plasmid to generate proteins that activate transcription on their own. As a consequence of this variability, we will score as putative interacting partners only those proteins that are identified in two independent screens, even though this criterion will result in the omission of some known interactions found only once (85).

Validate protein-protein interactions using affinity tags and MS-MS analysis. In addition to the two-hybrid system described above, we will also use recently developed tandem-affinity purification followed by mass spectrometry (29, 37) to determine protein-protein interactions. Isolation of protein complexes from a cell in native state followed by identification of the individual components by mass spectrometry has led to an alternative way of determining molecular machines. Availability of highly sensitive mass spectrometers such as Fourier transform ion cyclotron resonance–mass spectroscopy (FTICR-MS) has allowed detection and identification of proteins in subpicomole amounts, provided the genome has been sequenced. The limiting step in this method is availability of methods to purify and isolate protein complexes. A generic procedure to purify protein complexes in their native state was recently developed using tandem affinity purification (TAP) tags (70). This method allows isolation of protein complexes containing multiple components without the prior knowledge of composition or function of the molecular machine. The method involves attaching genetic tags to genes coding for selected proteins, allowing the modified proteins to be expressed in the native host, then using affinity columns to pull out the bait proteins and associated proteins. The individual proteins are extracted by a separation method such as SDS-PAGE, proteolyzed, and analyzed by MS/MS.

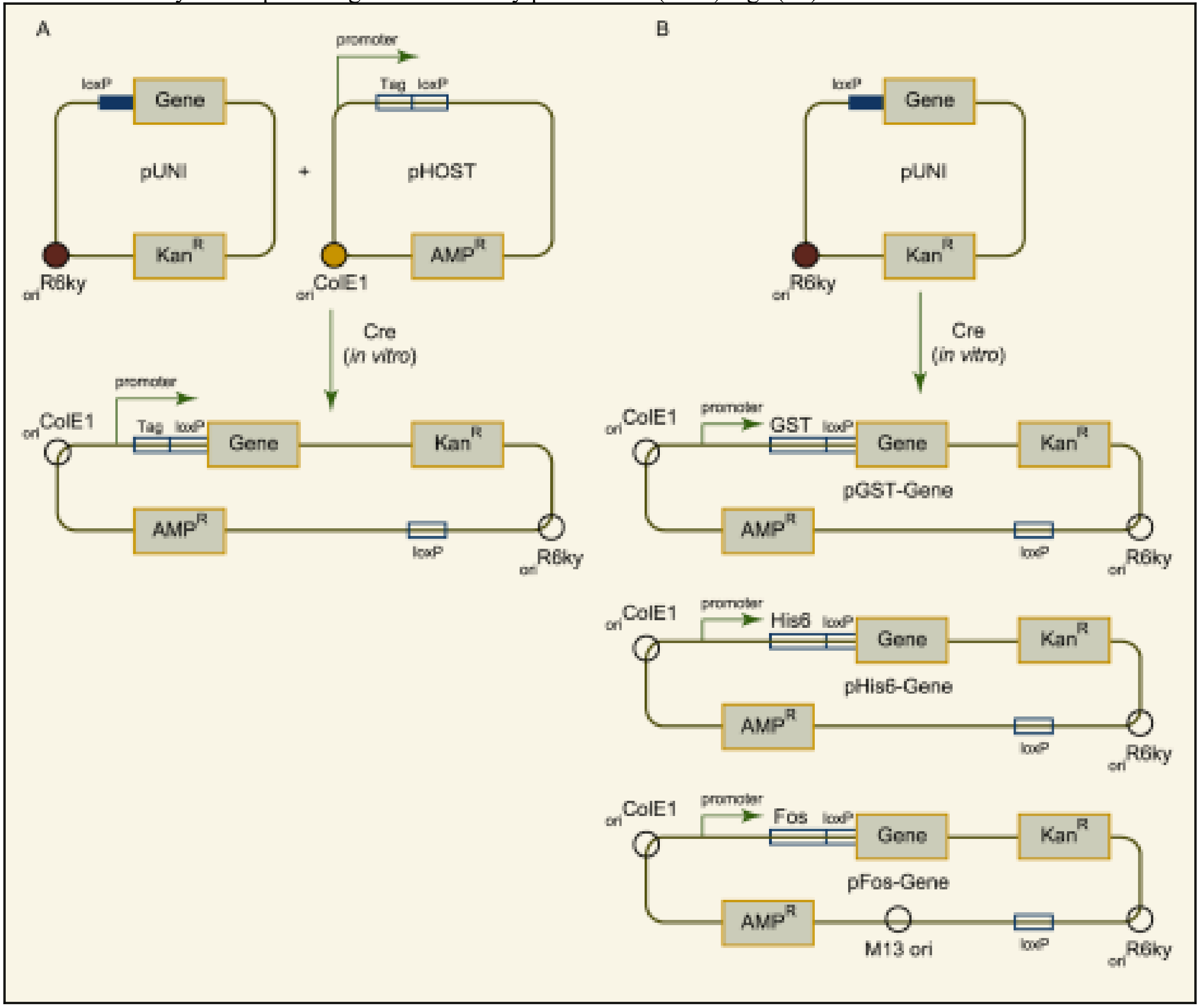


Figure 3.10. Univector plasmid fusion system. (A) Cre-*loxP* mediated site-specific recombination fuses the pUNI and pHOST plasmids at the *loxP* site. (B) As a result, the gene of interest is placed under the control of the pHOST promoter and fused to any Tag sequences present in the pHOST plasmid.

Protein complexes in *D. vulgaris*, *S. oneidensis*, and *G. metallireducens* will be purified and analyzed. *S. oneidensis* is a facultative anaerobe, whereas *D. vulgaris* and *G. metallireducens* are generally described as strict anaerobes. As it is evident that some protein complexes may be sensitive to oxygen, the isolation or cross-linking of complexes formed under anaerobic culture condition will be carried out in an anaerobic glove box.

Cassettes containing tags (poly-His or Protein A) will be inserted at the 3’ end of the genes encoding for the proteins central to the selected pathways. After selection of the positive clones, cells will be grown and collected in the midlog phase. They will be lysed mechanically with glass beads or by a cell homogenizer. Tandem affinity purification using low-pressure columns will be employed to “fish out” the bait protein and protein complexes with them. After separation by SDS-PAGE, the individual protein bands will be excised and either directly introduced to an FTICR-MS by electrospray or injected after digestion by a proteolytic enzyme such as trypsin. Repeating the experiments using several different bait proteins will identify the majority of proteins involved in the complex and ultimately in the pathway. The data generated using this approach will be complementary to the data generated using phage display.

3.4.4 Metabolite and Flux Profiling

Previously we have discussed how we will analyze the transcript and protein profiles of our target organisms. While the transcript profile indicates which genes are being expressed and the protein profile indicates the proteins present at a particular time and under a given set of environmental conditions, the physiology of the cell also depends on the processing of cellular metabolites into biomass. Just as there is no assurance that there will be a one-to-one correlation between the concentration of transcripts and corresponding proteins, there is no assurance that the presence of a given enzyme (as determined by protein profiling) will reflect the flux of metabolites through that particular enzyme (23). Thus, we need both metabolite and flux profiles of the cell if we want to characterize its physiology.

In this section, we will examine two types of profiles, the metabolite profile and the flux profile. The metabolite profile consists of the types and amounts of metabolites present in a cell at any give time under any given environmental condition. The metabolite profile is the steady-state (or pseudo-steady-state) pools of metabolites in the cell. In contrast, the flux profile is the rate of consumption or production of any metabolite (or through any enzyme). While the metabolite profile of a cell will give us an indication of the types of metabolites (and thus the types of enzymes) present in a cell, the flux profile will indicate which pathways are being used and at what amounts relative to each other. For example, using metabolite profiling techniques, one may be able to observe the presence of various pentose phosphate metabolites, which would indicate that the pentose phosphate pathway is functional in that particular organism; however, the metabolite profile would allow one to determine the relative amount of glucose-6-phosphate flowing through the pentose phosphate pathway versus glycolysis (12). A flux profile would yield that information and would lead to a better understanding of NADPH balances in the cell. Both are important for a detailed understanding of the cell. An overview of the sampling and analysis is shown in Figure 3.11.

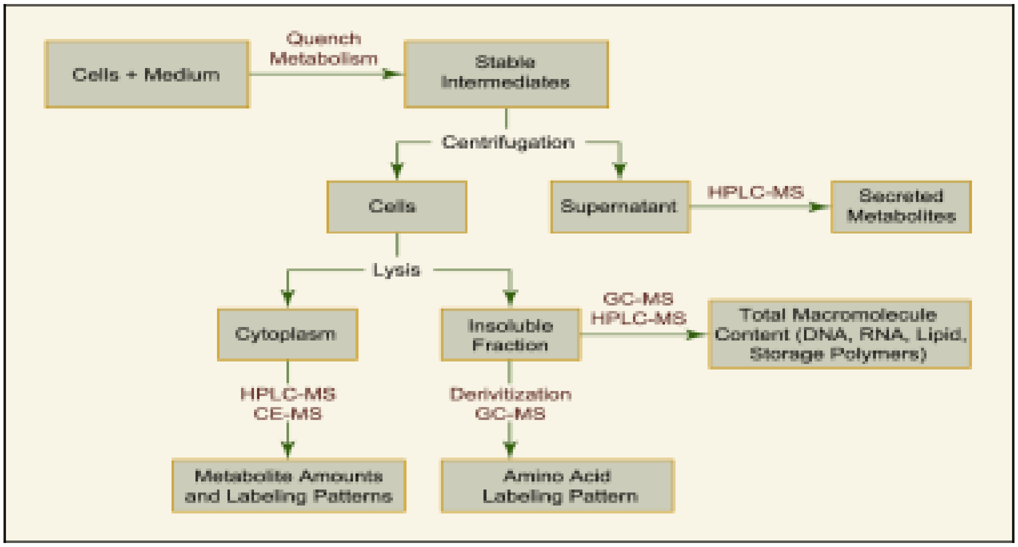


Figure 3.11. Flow diagram for metabolite and flux analysis of cells.

3.4.4.1. Cell Sampling. The methods used to sample the metabolites are extremely important, because of the rapid turnover of many metabolites (particularly ATP, ADP, etc.). A number of techniques have been developed for rapid sampling, most of which involve very short and rapid transfer of culture to a quenching liquid: liquid nitrogen, liquid CO_2, cold methanol, boiling ethanol, cold perchloric acid, or trichloroacetic acid (9, 46). The samples can then be centrifuged to separate cells from culture supernatant; the culture supernatant can then be analyzed for the utilization of carbon source and for the secretion of metabolites (e.g., waste products and secondary metabolites). The cells can then be extracted and intracellular metabolites and biomass constituents measured (see below). While it is not as important to perform rapid sampling when measuring the constituents of biomass (e.g., total amino acids, nucleic acids, lipids), one may want to fractionate the cells in such a way that intracellular metabolites and biomass constituents can be determined from the same cells. (When one is using stable isotope labeling to determine the distribution of isotopomers, the substrates can be very expensive.)

3.4.4.2. Cell lysis and separation. Once the medium has been separated from the cells by centrifugation, the cells will be lysed (pressure or sonication), and the insoluble fraction (macromolecules) will be separated from the cytoplasm by centrifugation. The content of the insoluble fraction will be determined using well-known and documented analytical techniques (30). For example, the fatty acid content of the membranes can be determined by making methyl esters of fatty acids and analyzing by GC-MS. We have shown that the fatty acid profile of cells can change rather dramatically with growth conditions (65). In a similar way, one can determine the total amounts of proteins and nucleic acids; while the amounts of these change significantly with growth rate/condition, the amino acid/nucleotide profiles do not change very much (65).

3.4.4.3. Flux profiling. Quantifying fluxes in fungi allows one to focus genetic manipulation on rate-limiting steps in a biosynthetic pathway. Flux-based metabolic models quantify the metabolic fluxes through all reactions included in an organism's metabolic network (61, 66, 80). By modeling the metabolism of a given organism, one can quantify the effect of genetic manipulations or changes to growth conditions on the cell's entire metabolic network. Inputs to flux-based models are the set of potentially active metabolic reactions and measurements of the steady-state production rates of metabolites such as DNA, RNA, chitin, carbohydrates, fatty acids, protein, etc. Improved estimates of the cell's fluxes can be obtained by feeding a ^{13}C-labeled carbon source and measuring the isotopic label state of resulting metabolites, such as amino acids or chitin (80). This type of approach is termed isotopomer analysis. This analysis will be directly supported by environmental simulator stress analysis in the Applied Environmental Microbiology Core (sub-section 2.4.4.8).

The network of reactions is represented by mole balances around each metabolite, which are contained in a stoichiometric matrix. Inverting this matrix and multiplying it by a vector of intracellular (zero) and extracellular (nonzero) net metabolite production rates gives the solution to the model: a vector containing flux values for each reaction (80). When isotopomer data are included, nonlinearities are introduced into the model, and iterative solution methods must be used (26, 27, 74, 80). Monte Carlo methods can be used to assess the sensitivity of the model to its inputs (26, 27, 74).

Nuclear magnetic resonance (NMR) spectrometry and mass spectrometry (MS) have been the most widely used techniques for determining the isotopic distribution in metabolites (17, 24, 93). In contrast to MS, NMR spectrometry gives the exact position of the isotopically labeled carbon in the metabolite of interest; however, MS is more sensitive than NMR, a characteristic important for many scarce intracellular metabolites (93).

To determine the flux in the central metabolic pathways, one can determine the isotopic labeling patterns in amino acids and carbohydrates. The labeling pattern in amino acids will be determined by hydrolyzing cellular proteins and derivatizing the resulting amino acids using either ethyl chloroformate or (N,N)-dimethylformamide dimethyl acetal (12). The derivatized amino acids will be analyzed by GC-MS. Labeling in central metabolites will be determined using HPLC equipped with ESI-MS-MS (9). The fluxes will then be determined using iterative solutions (26, 27, 74, 80, 92).

3.4.5 *Identification of Key Structural and Regulatory Components Using Targeted Mutagenesis*

One of the most powerful ways to define the function and regulation of a gene is to turn the gene off, or to change its normal pattern of expression by replacing the normal gene with a mutated counterpart. Once putative genes specific for cell adaptation are identified by microarray analysis as described above, we will generate deletion mutants defective in candidate stress-related genes in order to define their potential cellular functions. We will also create specific regulatory mutants to delineate the regulatory pathways and networks of stress-mediated response. We have successfully developed and optimized genetic vectors for the generation of deletion and insertion mutations in *S. oneidensis*. Several methods for generating insertion mutants have also been described for *D. vulgaris* and *G. sulfurreducens* (a very close relative of *G. metallireducens*).

3.4.5.1. Creation of Mutants

S. oneidensis mutants. Analyses of the annotated *S. oneidensis* genome have identified homologues of several conserved regulators (*oxyR*, *sodB*, *ompR*, etc.) implicated in cell adaptation and response to changing environmental conditions (Table 3.1). There are also a large number of genes encoding sensor proteins and response regulators (over 30 ORFs) (Table 3.2). Homologues of *phoB* and *phoR* genes, which encode a two-component system that regulates expression of the genes induced by carbon, nitrogen, and phosphate starvation in *E. coli* (41), have also been identified in the *S. oneidensis* MR-1 genome.

As previously shown for *E. coli* (81), synthesis of stress-related proteins is subject to hierarchical regulation by various environmental factors. The regulatory genes, either induced by specific stress conditions or involved in global cellular responses, will be chosen as primary targets for gene replacement. Based on the experimental results and feasibility, mutants defective in a number of regulatory genes will be generated. Since many functionally unknown ORFs could be identified using DNA arrays, we will prioritize these ORFs for mutagenesis based on their expression levels, expression conditions, and putative functions.

Several allele replacement techniques, which include in-frame deletion mutagenesis, have been developed and used successfully at ORNL. The PCR-based in-frame deletion will be performed using a modification of a method described by Link et al. (49) (Figure 3.12). The mutated versions of the target genes will be generated by crossover PCR, where a pair of "outer" primers will be designed to amplify the gene of interest plus 1 kb of flanking DNA on each end of the gene. The designed primers will carry a 5'restriction site containing a central CTAG sequence (i.e., SpeI, XbaI, NheI, or AvrII). These sites are relatively rare in *S. oneidensis* MR-1 genes, and several exist as unique sites in plasmids used for cloning in these experiments. A second set of "inner" primers will be designed to overlap the N-terminal or C-terminal and seven codons of the gene separately, and to amplify this with the attached flanking DNA using the appropriate "outer" primer. The inner primers will each carry a 21 bp linker sequence (TATTTAAATTTAGTGGATGGG) at their 5' end such that the resulting PCR products will contain a 21 bp overlap. A final round of PCR using these two separate but overlapping PCR products and the "outer" primers will generate a PCR product containing the flanking DNA sequences and a deletion of nearly the entire gene of interest. The deleted region will be replaced by the 21 bp linker sequence, retaining the original reading frame and introducing a SwaI restriction site into which antibiotic resistance cassettes can be cloned to make polar mutations. The mutated versions of the DNA fragment will then be cloned into pDS31 (Beliaev et al., unpublished results), an R6K suicide vector that carries *sacB* and *aacC1* (Gm^r) cassettes, and possesses a □ protein-origin of replication. The recombinant plasmid will be introduced into a Rif^r strain of MR-1 by conjugation. The integration of the suicide construct into the chromosome will be selected by Gm/Rif resistance and will be mapped by PCR and sequencing. After the integration of the suicide plasmid is confirmed, one of the resulting mutants will be grown in broth without Gm and will be plated onto media containing 5% sucrose (7). Since the presence of sacB gene confers sucrose sensitivity, we will select cells that have undergone a second round of homologous recombination and have resolved the suicide plasmid. During this event, the cell will either revert to the wild-type allele or will retain the mutated allele. Thus, a fraction of the resulting sucrose-resistant colonies will carry a deletion mutation in the target gene. Those mutants will be screened and selected by colony PCR using primers flanking the deleted region.

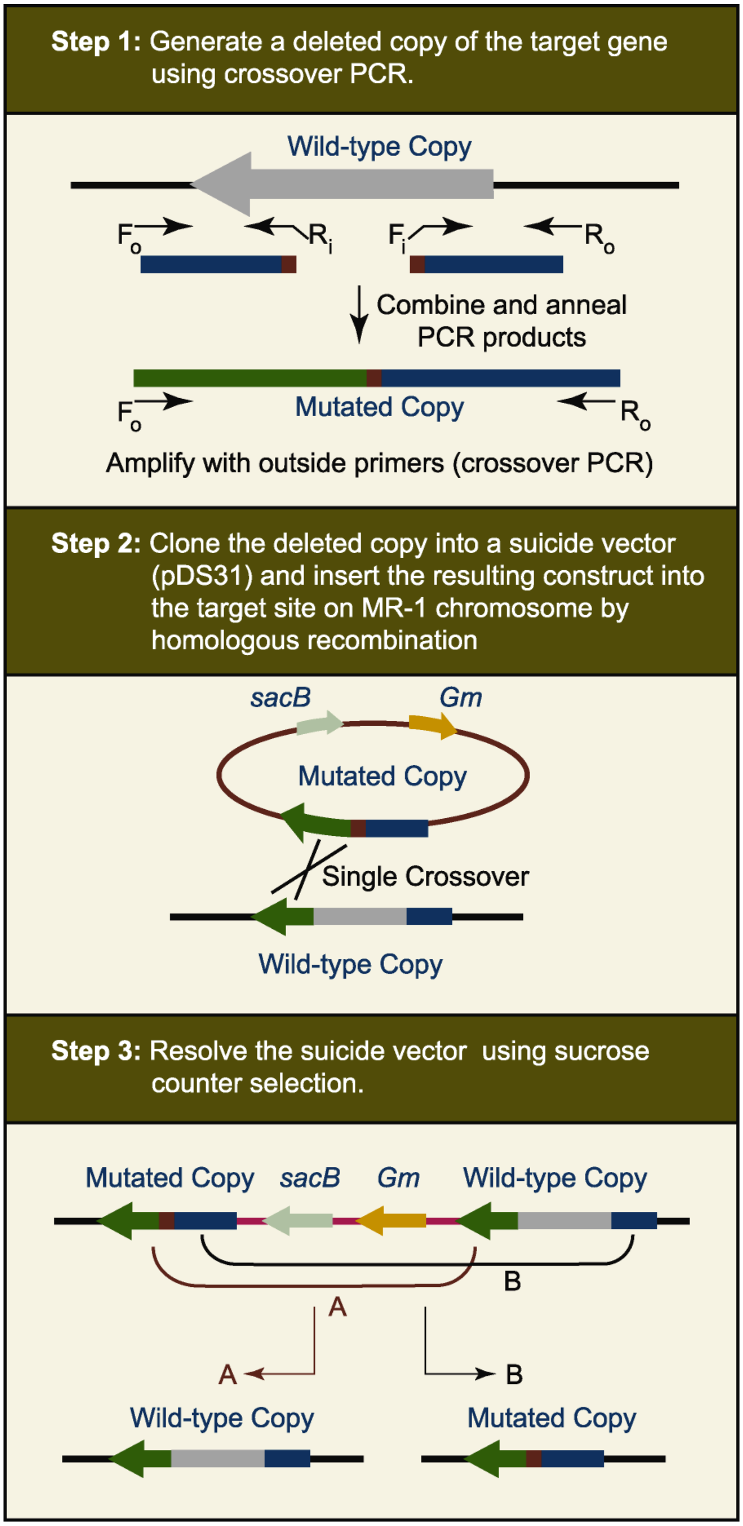

Step 1: Generate a deleted copy of the target gene using crossover PCR.
Wild-type Copy
Fo
Ri
Fi
Ro
Combine and anneal PCR products
Fo
Mutated Copy
Ro
Amplify with outside primers (crossover PCR)
Step 2: Clone the deleted copy into a suicide vector (pDS31) and insert the resulting construct into the target site on MR-1 chromosome by homologous recombination
sacB
Gm
Mutated Copy
Single Crossover
Wild-type Copy
Step 3: Resolve the suicide vector using sucrose counter selection.
Mutated Copy
sacB
Gm
Wild-type Copy
B
A
A
B
Wild-type Copy
Mutated Copy

Figure 3.12. Schematic representation of the in-frame deletion mutagenesis procedure for *S. oneidensis* MR-1.

In addition to single mtutants, we will generate mutations in multiple regulatory genes implicated in stress-induced response. *S. oneidensis* mutants defective in multiple regulatory stress-related genes will be very important for dissecting the stress-regulated regulatory networks.

We will complement each obtained mutant using the broad-host-range vector pRK415 (42). The complete sequence of the target gene(s) will be amplified by PCR using primers complementary to the upstream and downstream regions. The gene(s) will be cloned into the pRK415 multiple cloning site, which is located downstream of *lac* promoter. In order to allow expression from the *lac* promoter, we will attach 6 bp restriction sites at the 5'-end of each oligonucleotide primer for directional cloning. The recombinant plasmid(s) will be transferred into the mutant(s) by conjugation. The complemented strains will be examined as described above.

***D. vulgaris* Hildenborough.** Gene replacement has been accomplished in *D. vulgaris* by Voordouw and coworkers (22, 28, 62, 86). Their approach has been to make a marker exchange plasmid by placing DNA from upstream and downstream of the gene of interest on either side of an antibiotic resistance gene. This construct is then introduced by conjugation into the strain to be mutated, and a single recombinational event that inserts the entire plasmid into the chromosome is confirmed. Then a second recombinational event that generates the gene deletion is signaled by the absence of both the plasmid-borne antibiotic resistance and levan sucrase gene *sacBR,* which are enriched by sucrose resistance. This is certainly a functional approach, but mutants are not generated rapidly.

Currently kanamycin and chloramphenicol resistances are the only well-characterized antibiotic resistances useful for *D. vulgaris* selections.

Improvements are needed to increase the rate and number of mutants obtained through molecular manipulation. A procedure for the construction of mutagenic plasmids used by Maier and Myers (53) for *Shewanella* will be used for most deletion constructs. The approach is to amplify the gene of interest plus some flanking sequence, about 1500 bp, by PCR. The PCR fragment is captured in a plasmid such as pGEM®-T Easy (Promega Corp., Madison, WI). The gene deletion is created by a second set of primers separated by ca. 250 bp reading out of the gene of interest used to drive inverse PCR (IPCR). This IPCR fragment is ligated to an antibiotic cassette. If electroporation is feasible, this plasmid, which should not replicate in the bacterium to be mutagenized, could then be introduced with selection for the antibiotic retention and screened for the loss of plasmid sequences. If conjugation is necessary to introduce the DNA into the cell to be mutagenized, the original primers can be used to amplify the chromosomal DNA fragment with the antibiotic resistance gene replacing a portion of the target gene. The amplified fragment can then be put into a conjugable vector for subsequent mutant generation (53).

Disruption of the restriction-modification system. Unlike *Geobacter sulfurreducens*, which is readily electrotransformed (14), the introduction of recombinant DNA into *D. vulgaris* by electroporation does not occur at a workable efficiency. This difficulty has been reported informally by several researchers and has been noted by us. In addition, few recombinants are generated following the introduction of DNA by conjugation. The genome sequence reveals the presence of several (two or three) restriction modification systems that surely limit recombination events. Disruption of one or more of the restriction systems should increase the efficiency of mutant construction in *D. vulgaris.* An internal fragment of a type II restriction enzyme that is a possible ortholog to *PaeR*7I has been cloned and will be used to disrupt the gene to test the effect on recombination efficiency.

If a positive result is obtained, the next step will be to create a deletion of this restriction endonuclease by marker exchange. We propose to develop a system adapted from Datsenko and Wanner (16) to eliminate the antibiotic resistance marker once the deletion has occurred. Directly repeated FLP recognition target (FRT) sites will flank the antibiotic resistance marker replacing the gene for the restriction enzyme. Once the double recombination has occurred, the antibiotic resistance will be eliminated by the introduction of the gene for the FLP recombinase (2). The recombinase gene will be introduced on a replicon that can be cured from the cells by increasing the temperature. For the SRB, a temperature-sensitive replicon will be generated from the cryptic plasmid, pBG1 (88), that has been used for shuttle vector construction. The resulting construct will carry a scar of the FRT motif at the site of the deletion of the restriction enzyme gene but none of the antibiotic resistance gene or other plasmid sequences. The antibiotic resistance used to create this deletion can then be used again to generate mutations in a number of regulatory genes. This FLP/FRT system should expand the number of mutations that can be generated by marker exchange, but the FRT motif scar will make multiple use of the system, though it will require careful monitoring.

Transformation with linear dsDNA. The ability to transform a bacterium with linear dsDNA increases the efficiency of recovering the double recombination event necessary to replace a wild-type gene with a deleted, marked copy. A single recombination with a linear DNA substrate results in a double-strand break in the chromosome, which, if not repaired, can be lethal. Thus there is a strong selection for the double recombination event that restores the integrity of the chromosome. Linear dsDNA, however, is a good substrate for exonucleases and does not generally last long enough to promote recombination events in wild-type cells (51). In *E. coli* and several other gram-negative bacteria, the very active nuclease ExoV that is a part of the RecBCD complex invades dsDNA ends and rapidly degrades the DNA. To replicate its DNA in its *E. coli* host by a rolling-circle mechanism, bacteriophage lambda encodes a protein that inhibits the activity of ExoV. Does such an exonuclease system limit genetic exchange in *Desulfovibrio*? Can we use the lambda inhibitor to increase the frequency of recombination?

Sequence gazing of the *D. vulgaris* genome reveals an ortholog to the *recBCD* exonuclease system of *E. coli* present in the genome. We will test the effectiveness of the introduction of the Red system of the bacteriophage lambda for increasing observable recombination events (16, 55). Three genes are a part of this system: □-encoding Gam, an inhibitor of the exonuclease V activity of RecBCD; and □ and exo, which form a recombinase that can insert linear DNA into a homologous chromosome. If Gam can interact with the SRB exonuclease V enzyme, we would predict an increase in recombination even in the presence of functional restriction enzymes. If Gam does not inhibit the activity of the exonuclease V of the SRB, providing an additional recombinase to the cells may allow recombination to be detected from the introduction of linear fragments of DNA. These genes will be introduced into

D. vulgaris and monitored by RT-PCR for expression. Several SRB promoters are available if transcription appears to be limited. The Red genes can be placed in single copy on the chromosome in a neutral site at the Tn*7* attachment site (3, 89). The Red SRB host will then be used as a recipient for marker exchange events with linearized DNA fragments.

Surprisingly, orthologs of two genes, *comEC* and *comF,* reported to be involved in the acquisition of transformation competence in gram-positive bacteria, are recognized in the genome of *D. vulgaris.* Orthologs of these genes also occur in the naturally transformable *Hemophilus influenzae* (Ecocyc.org). The possibility that *D. vulgaris* could develop natural competence during growth has not been assessed carefully. To test this possibility, DNA from a nalidixic acid-resistant mutant of *D. vulgaris* will be prepared and added to samples of cells of a sensitive strain. The recipient cells will be prepared at timed intervals throughout the growth cycle of the bacteria. It is possible that species-specific sequences are necessary for the DNA uptake, as in *Neisseriae* (76). Alternatively, an amino acid auxotroph of *D. vulgaris* can be isolated from the deletion constructs and used as the recipient for wild-type DNA with selection for prototrophy. A positive result would be a welcome surprise.

Transposon mutagenesis. Transposon mutagenesis has been applied successfully in many bacteria to create mutations in many genes in a genome quite rapidly. Because of the low efficiency of introduction of DNA by electroporation, as well as conjugation, coupled with an apparently (word missing?) transposition frequency, this approach has not been used routinely with the three bacteria targeted in this proposal. If we are successful in improving the efficiency of electrotransformation or conjugation by the procedures above, we will again attempt to generate a library of transposon mutations in *D. vulgaris.* The vectors to be used will be the mini-Tn*5* transposon derivatives, pUT/Km, pUT/Sm/Sp, pUT/Cm, and pCNB5 Tc^{R} (18, 19, 36). These vectors have the advantage of being suicide plasmids in the SRB. The transposase is not part of the transposon inserted into the chromosome, and the transposase gene is lost along with the plasmid after transposition. Therefore the minitransposon should be quite stable. To improve the number of transposition events, expression of the transposase can be increased by the introduction of a strong SRB promoter, as demonstrated by Krumholz and coworkers (personal communication). pCNB5 Tc^{R} also creates nonpolar mutations so that the interpretation of mutant phenotypes can be more straightforward. We propose to isolate a library of about 10,000 mutational events. Chromosomal sequencing or sequencing of PCR products generated from the transposon will allow rapid identification of the insertion sites. Of primary importance for the creation of this library of mutations will be the growth conditions of the cells during the isolation. Cells growing with different terminal electron donors and/or acceptors will be mutated to allow different electron pathways to be targeted. Those mutants generated by transposon insertion that have altered responses to stress will be those first targeted for sequencing.

Regulated promoters. Genes that are essential under all growth conditions established in the lab will not be identified among those mutated. To evaluate the effects of disruption of essential functions, regulated promoters are also necessary. By cloning a wild-type copy of the gene on a stably maintained plasmid, or by introducing a duplicate of the gene into the chromosome in single copy (3, 89), the original chromosomal copy can be disrupted while the essential gene product is supplied from the plasmid. A regulated promoter could then allow the plasmid-borne gene to be shut off and the resulting phenotype to be examined. Candidate promoters include the well-studied lac promoter. Many lac-promoter expression vectors with a single operator have been reported to be leaky. However, a more complete understanding of the operator region has shown that the inclusion of an auxiliary operator provides for extremely tight regulation by the *lac* repressor (90). This system can be adapted for use but depends on the ability of the cells to transport the gratuitous inducer. Removal of the inducer and subsequent decay of mRNA and preexisting protein add uncertainly to an interpretation of global mRNA levels but might allow physiological observations of importance.

A second system that is currently being examined for a regulated promoter in the SRB is the tetracycline resistance gene promoter. TetA promoter is stringently repressed until chemical induction in *E. coli* with about 200 ng/ml anhydrotetracycline (described for the *Strep*-tag□ Expression System, IBA GmbH, Gottingen, Germany). We have determined that the minimum inhibitory concentration (MIC) for tetracycline for SRB is about 3 □g/ml, well above the concentration needed to induce the promoter construct, and we will be testing this promoter for control in the SRB in the near future.

Key biochemical and regulatory functions implicated in stress responses and metal reduction will be targeted for mutation. Although having a better genetic system for *D. vulgaris* will improve the speed of mutant generation, the ability to obtain mutants in a number of genes that allow responses to stress does not depend on those developments. We will initiate the creation of a battery of mutations in alternate sigma factors and orthologs of genes encoding *E. coli* DNA-binding proteins (71). These mutants will provide comparison genetic backgrounds when regulatory circuitry is being elucidated.

Putative sigma factors whose genes are recognized in the *D. vulgaris* genome are RpoD, RpoN, RpoH, and FliA. The latter three control nitrogen-regulated promoters, heat-shock responsive promoters, and promoters for flagellum biosynthesis, respectively. Deletions of these genes should be possible.

Those genes encoding DNA-binding proteins for which the binding sites have been identified by footprinting in *E. coli* will also be targets for mutagenesis. Of the fourteen that apparently have recognized orthologs in *D. vulgaris*, those encoding Fnr-like proteins, Dnr and HbaR, the Fur protein possibly regulating ferric uptake, PhoB regulating the phosphate regulon, and ModE for molybdenum metabolism, will receive attention first. As data accumulate from the microarray experiments and from modeling predictions, new targets will be identified for confirmation of the models of regulation.

Exposure to oxygen is certainly a serious stress for bacteria known to be killed by atmospheric levels. Since bioremediation environments are often oxygenic, the response of metal-reducing bacteria to this added stress needs investigation. The ability of the classically "strictly anaerobic" *Desulfovibrio* strains to handle oxygen has been well established in the last two decades (21, 34, 40, 73). Still, these bacteria are sensitive to atmospheric concentrations of oxygen and have several mechanisms for responding to this stress, including chemotaxis (28), superoxide reductase (15), and respiration (47, 87). Mutants in the genes encoding several of these systems have been created to confirm the suggested functions. These mutants will be requested for analysis here. If not available, new mutations can be constructed following the published procedures.

We have identified the MIC of uranium for the sulfate-reducing bacterium *Desulfovibrio desulfuricans* strain G20. Obtaining this information for *D. vulgaris,* we will isolate mutants that are more resistant to this toxic metal. By complementation with a plasmid library of DNA fragments of the mutant DNA, we will attempt to isolate the gene responsible for the increased resistance. This gene should indicate one of the target sites for uranium inhibition in the cell and could provide insight into the stress response of cells exposed to toxic metals. This protocol assumes that resistance will be dominant. We will also screen the transposon mutants to determine whether any have an increased resistance to metals.

Care will be taken to subculture any mutants the fewest number of times possible before storage at -80^{-}C. It is clear that compensatory mutations accumulate in regulatory mutants with time, as was observed in microarray analyses of TyrR mutant strains of *E. coli* (43). Physiological characterization of the mutants will include rates of growth with, and utilization of, various electron acceptors (sulfate, sulfite, thiosulfate, fumarate, dimethyl sulfoxide, uranium, chromium, etc.) and electron donors. These experiments will be necessary to establish the preliminary conditions for preparation of mRNA for the microarray experiments. Biochemical characterization will complement the physiology and sequencing information.

As natural populations are sampled for the diversity of microbes in the environment of our target organisms, new isolates may be obtained. Any isolates of *Desulfovibrio* will be screened for increased resistances to the stresses under study. The genes conferring the increased stress will be sought by molecular approaches. A plasmid library of the chromosomal DNA will be prepared and transferred into *D. vulgaris* with selection for increased resistance. Any genes that improve the stress response of *D. vulgaris* will be identified and compared to the endogenous ortholog to determine the critical features for resistance.

***G. metallireducens* mutants.** Lovley and colleagues have developed techniques to create *G. sulfurreducens* mutants using a gene-replacement system (14). In brief, *G. sulfurreducens* is transformed with a linearized, suicide plasmid containing regions contiguous to the gene of interest and an antibiotic resistance cassette between the contiguous regions. Lovley and colleagues have found that double recombination occurs at a relatively high frequency. Given the similarities between *G. metallireducens* and *G. sulfurreducens*, we anticipate that a similar system can be used to interrupt genes in *G. metallireducens*. We will use similar techniques to eliminate genes thought to be involved in the various stress responses.

3.4.5.2. Characterize Phenotypes Exhibited By Mutants Defective in Key stress-related Regulatory or Structural Genes. The resulting regulatory and structural mutants will be analyzed for their ability to adapt to different environmental stresses, as well as to utilize various electron acceptors. The mutant strains will be tested for their ability to reduce various metals under varying pH, phosphate availability, salinity, metal concentrations (e.g., Cd, Hg, Ni, and Cr), and nutrient concentrations. The metal electron acceptors will include different forms of Fe(III), Mn(IV), and U(VI). Both wild-type and mutant strains will initially be grown to an early stationary phase. These cells will then be transferred to the appropriate medium containing the appropriate electron acceptor. The electron-acceptor reduction rates in the wild type and the mutants will be measured spectrophotometrically as described previously (56).

At the same time, these strains will be tested for growth in liquid culture under oxic (*Shewanella* only) and anoxic conditions with the following nonmetal electron acceptors: nitrate, nitrite, fumarate, trimethylamine N-oxide, dimethyl sulfoxide, tetrathionate, thiosulfate, and sulfite. Growth under these conditions will be followed spectrophotometrically to allow quantitative comparisons of growth rates using these electron acceptors. The reduction of nitrate, nitrite, thiosulfate, and sulfur will be determined as described elsewhere (4). Reduction of and growth on elemental sulfur as an electron acceptor will be examined using solid media, as previously described (54).

To confirm the role of the mutated ORF in the determined phenotype, the wild-type ORF expressed from its own promoter will be introduced into the appropriate mutant strain on a low-copy-number plasmid pRK415 (42). These strains will be examined as above for their electron-acceptor utilization and reduction phenotype.

3.4.5.3. Elucidate the Genetic Basis Underlying Regulatory Mechanisms of Stress Response Using Transcriptomics, Proteomics, and Metabolomics. The changes in gene expression patterns in response to environmental stress are complex, with the expression profiles of many genes being altered. One approach to defining the contribution of a given regulatory gene to a complex cellular process is to use DNA microarrays to monitor the gene expression patterns of the mutants defective in individual regulatory genes (20). In this task, we will use transcriptomics, proteomics, and metabolomics to study the expression profiles in the regulatory mutants of *D. vulgaris*, *S. oneidensis*, and *G. metallireducens* generated as described above.

Both wild-type and mutant strains will be grown as described above. RNA extraction, cDNA synthesis and labeling, and microarray hybridization and scanning will be performed using standard protocols. The cDNA prepared for wild-type and mutant cells will be labeled with Cy3 and Cy5, respectively, mixed together in equal amounts, and hybridized to the microarrays. Proteins and metabolites will be analyzed in a similar manner.

By comparing the transcript, protein, and metabolite profiles between wild-type and mutant strains defective in one or more regulatory genes, we should be able to understand which genes are repressed or activated by the putative regulatory gene under the conditions of the experiment. Therefore, by comparing gene expression patterns between wild-type and regulatory mutants, we should be able to evaluate the contribution of individual regulatory genes to stress-induced cellular responses and identify the genes controlled by these regulators.

3.4.6 *Disruption of Signaling/Metabolic Pathways Using Chemicals*

Another way to interrogate the interactions between proteins in the stress response or to interrupt key metabolic pathways that might be involved in the stress response is to add chemical inhibitors. The goal of this portion of the proposed work is to develop molecular inhibitors for key members of the stress response networks using the combinatorial-chemistry approach of small-molecule libraries combined with screening against selected targets. Starting with a small number of selected targets, the ultimate goal will be to identify a chemical inhibitor for every interaction in a pathway.

First, we will screen large combinatorial libraries of small molecules using the phage display and yeast two-hybrid systems described above. Once we have identified a series of compounds that inhibit protein-protein interactions using these assays, the compounds will be screened for their effect on the response of the organism to the particular stress that the target proteins are known to inhibit. In this test, the organisms will be first grown (non-stress) in the presence of the compound to see if it affects growth. Next, the compound will be added at various times prior to the application of a particular stress. The response of the organism to the stress (in the presence of potential stress response inhibitors) will be monitored using one or more of the physiological profiling techniques described above. This test will further narrow the list of potential inhibitors as some of these will be unable to penetrate the membrane. Finally, the narrowed list of inhibitors will be tested in conjunction with various mutants in the stress response pathways.

3.4.7 BioPanning for Environmental Stress Response Pathways

Diversa Corporation will apply a combination of advanced technologies, including large DNA insert cloning, biopanning, and ultra-high-throughput (UHTP) screening, to study the distribution of stress response pathways in the environment. Diversa routinely constructs large insert fosmid libraries from environmental samples collected worldwide. We will construct large insert fosmid libraries from samples obtained from our collaborators. These will be screened using biopanning, Diversa's patented technology for enriching desired sequences from DNA libraries by solution-phase hybridization. In this instance, biopanning will involve Gel MicroDroplet (GMD) in situ hybridization followed by UHTP screening with FACS (Fluorescence Activated Cell Sorting). This approach will be used to capture *E. coli* fosmid clones carrying genes encoding enzymes responsible for stress response.

Estimates of cultured microorganisms vary from 0.001–0.1% in seawater, 0.25% in fresh water, 0.3% in soil, to 1.0–15% in activated sludge (1). Therefore, in order to explore the distribution of stress response pathways in the environment, a culture-independent approach is required. Diversa Corporation has pioneered such an approach to explore this untapped resource and is now a global leader in the recovery of DNA directly from the environment and in the production of large insert DNA libraries. We have constructed proprietary microbial DNA libraries from a multitude of terrestrial ecosystems (including soil, activated sludge, deep ocean vents, Antarctic rocks, geothermal pools, plant roots, and insect guts) in cloning vectors that can be propagated in *E. coli*, *Streptomyces*, and other appropriate hosts. In order to facilitate discovery of stress response pathways in this extensive amount of untapped microbial genomic material, an ultra-high-throughput (UHTP) DNA screening approach is necessary. The project will exploit Diversa's considerable strength and expertise in the areas of DNA extraction, library construction, and FACS.

3.4.7.1. Large Fragment Cloning. Since stress response pathways are clustered on chromosomal DNA fragments and generally vary in length from 20–40 kb, it is essential to clone large DNA fragments to capture entire pathways. We have overcome two hurdles to successfully clone large genomic fragments from the environment: (1) the low cloning efficiency of environmental DNA, and (2) the inherent instability of large DNA clones. Diversa scientists have developed effective DNA extraction methods and vector/host systems that allow stable propagation of large DNA fragments in *E. coli*. They have also developed methods that allow estimations of both the extent and nature of the diversity present within these environmental libraries.

Processed environmental samples are embedded in agarose noodles for protein digestion and release of high molecular weight (>100 kb) DNA. DNA is electroeluted, partially digested with restriction enzymes, and size-selected by agarose gel electrophoresis. It is then ligated to fosmid arms and packaged into phage lambda particles that are used to infect *E. coli*. Clones are then arrayed in microtiter plates. The microbial diversity of the libraries is determined with Terminal Restriction Fragment Polymorphism (T-RLFP) (13). *The E. coli* F factor replicon (44) contained on our vector maintains the fosmid at 1–2 copies per cell in *E. coli*. The low copy number is essential for stably maintaining large environmental DNA inserts.

3.4.7.2. Large Insert FACS Biopanning. The individual *E. coli* clones of our large insert libraries will be encapsulated in GMDs and incubated to permit clonal amplification. Clones in GMDs will be hybridized in situ with oligonucleotide probes targeting conserved regions of stress response pathways. Hybridized GMDs will be sorted

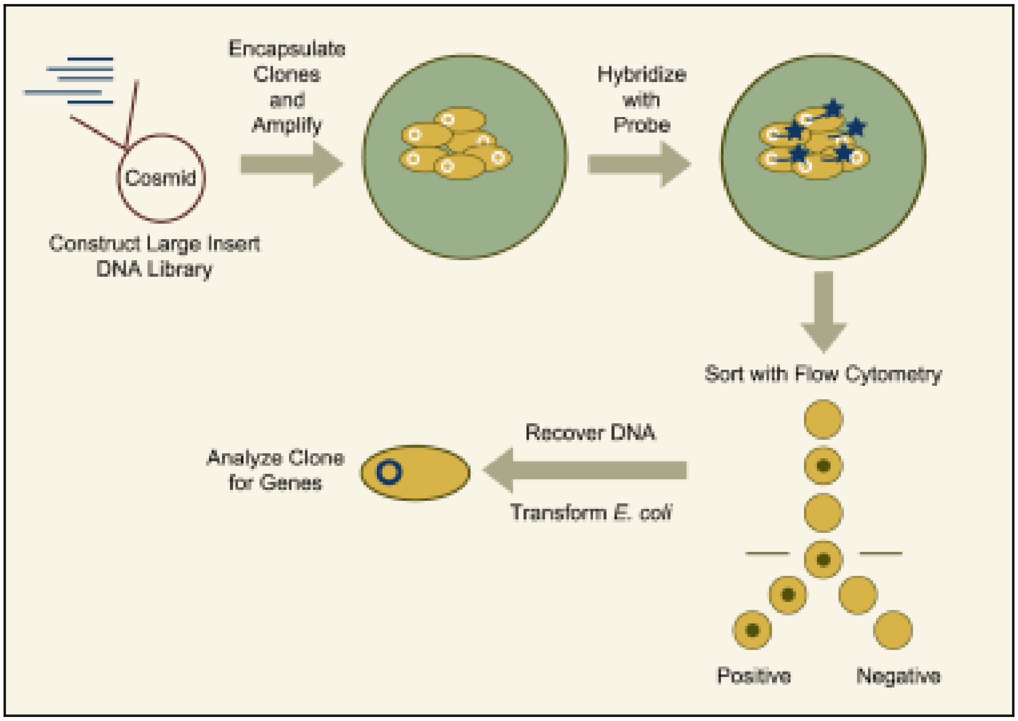


Figure 3.13. Large insert cloning experimental design.

by FACS and collected. This approach of using GMDs for clonal enrichment and hybridization followed by FACS is called “large insert FACS biopanning.” The DNA in the sorted GMDs will be extracted and introduced back into *E. coli* by transformation or electroporation. This experimental design is summarized in Figure 3.13.

The high-capacity representation of clone diversity available with GMD technology, combined with the rapid screening capability of Diversa’s FACS technology, enables the UHTP screening capability of our approach. The GMD In Situ Hybridization (MISH) allows the rapid enrichment and isolation of rare-pathway genes (58, 64). Five hundred microliters of 3% agarose can yield up to 10^6 GMDs, and under our encapsulation conditions, approximately 10–20% of the GMDs will be occupied by 1–2 cells, so theoretically, 10^5 library clones can be screened in one experiment. The technology is facile and can be readily adapted to automation. Diversa has three proprietary FACS machines in operation, screening GMDs at a rate of 5,000/sec and yielding a maximum theoretical throughput of 10^8 GMDs per day. With a 10% occupation rate, this translates to 10^7 clones per day. GMDs will be made by dissolving 3% (w/v) Seaplaque GTG low-melting temperature agarose (BioWhittaker Molecular Applications, ME) in phosphate-buffered saline (PBS). The percentage of agarose dictates the pore size of the GMD and has been optimized in our preliminary experiments to allow free penetration of nutrients, antibiotics, nucleic acid probes, and antibodies, which are the components of MISH (64). According to the manufacturer, up to 300 kD molecules will diffuse freely into 3% GTG agarose. This percentage also will stabilize the GMD during alkaline denaturation and in hybridization conditions of up to 46°C in 50% formamide. Exponentially growing *E. coli* cells containing the large insert library will be mixed with 45°C molten agarose by vortexing at full speed, followed by the addition of mineral oil. The mixture will be emulsified with a CellSys100 Microdrop maker (licensed from One Cell Systems). Optical density and the volume of *E. coli* cells mixed with molten agarose have been optimized to achieve 1–2 cell occupancy/GMD. The GMDs will be washed twice with PBS and incubated in growth media under the appropriate conditions to achieve large microcolonies without allowing the cells to grow out of the GMDs. The GMDs will be harvested by centrifugation and subsequently used for *in* situ hybridization.

3.4.7.3. GMD In Situ Hybridization (MISH). Microcolonies contained in the GMDs will be lysed and deproteinized to expose their DNA for hybridization. Denatured digoxigenin (DIG)-probes will be hybridized to the DNA in the GMD’s. The fluorescent signal for FACS sorting will be generated and amplified by tyramide signal amplification (TSA), an enzyme-mediated signal-amplification method (8). The increased sensitivity offered by TSA has been shown to be important for detection of short oligonucleotide probes and low-copy genes by fluorescent in situ hybridization (79).

3.4.7.4. Sorting by FACS and Recovery of DNA from GMD for Transformation. Hybridized GMDs will be sorted by FACS using a fluorescein filter at an excitation of 488 nm and an emission of 510–550 nm. Occupied GMDs will be further differentiated based on fluorescence intensity. Those occupied GMDs with highly fluorescent signals will be sorted into 96-well plates for recovery of DNA. The GMDs will then be melted and digested with agarase (New England Biolabs) to remove agarose, dialyzed to remove the fluorophore and other impurities, and then introduced into competent *E. coli* hosts by electroporation or transformation.

3.5 Experimental Core Facilities

3.5.1 Metabolite/Protein Profiling Facility (LBNL, U.C. Berkeley)

We have requested funds for equipment to profile metabolites and proteins. This facility will be equipped with devices for the separation of metabolites and proteins using chromatography techniques and identification of the metabolites/proteins using mass spectrometry (MS) detection. While the facility will concentrate its efforts primarily on profiling a large fraction of the metabolites in a cell (analogous to transcript profiling with DNA arrays), this same equipment may also be used to separate and detect proteins (protein profiling). This facility will be equipped with the following: (1) Gas chromatographs (GC) equipped with mass spectrometry (MS) and flame ionization (FI) detectors will be used for measuring fatty acids, volatile secondary metabolites excreted into the culture medium, and derivatized metabolites and amino acids. By using a [^{13}C]-labeled substrate, derivitizing metabolites and amino acids, and mass spectrometry detection, we can perform isotopomer analysis to determine fluxes. (2) High performance liquid chromatography (HPLC) pumps equipped with UV-Vis, MS, and flo-thru radioisotope detectors for separating intracellular metabolites, extracellular (excreted) metabolites, etc. MS detectors will be used for identifying metabolites that cannot be detected by another means and for measuring isotopomer balances. (3) Capillary electrophoresis (CE) equipped with UV-Vis and MS detectors for separating charged metabolites. CE has been shown to separate greater than 100 metabolites simultaneously. (4) The MS detectors for HPLC and CE will be electrospray injection-quadrupole (ESI-MS), ESI-quadrupole-quadrupole (ESI-MS-MS), and ESI-quadrupole-time of flight (ESI-QTOF) spectrometers. (5) Robotics equipment will be used for high-throughput processing of samples. The robotics equipment will include liquid handling devices and autosamplers to automate as much liquid handling as possible. We will purchase some duplicate pieces of the most heavily used equipment in later years to make the facility a full-fledged user facility. In addition to the equipment, we have requested funding for a full-time technician with expertise in mass spectrometry and chromatographic separations. This person will be primarily responsible for maintaining equipment, training users in the equipment operations, and developing new methodologies.

In addition to the equipment offered in the metabolite/protein profiling facility, users will have access to state-of-the-art analytical equipment in the College of Chemistry at the University of California at Berkeley. This includes 400-MHz and 500-MHz nuclear magnetic resonance (NMR) spectrometers. These may be particularly useful in confirming the identity of a particular metabolite.

3.5.2 Combinatorial Chemistry for Functional Genomics Facility (LBNL, U.C. Berkeley)

We are in the process of building a core facility at LBNL for the implementation of small-molecule based functional genomics that integrates the following central components: (1) a collection of vastly diversified and highly effective small molecule libraries, (2) a group of novel high-throughput biological assays, (3) a professional team maintaining facility's efficient operation, and (4) development of advanced detection method and instrumentation. Such a facility would provide the compounds, assays, functional deconvolution strategies, robotics, state-of-the-art detection instruments, data processing/management software and technical support so that inhibitors can be quickly and cost-effectively developed.

The facility will leverage the considerable expertise of chemists from UC Berkeley who are the world's experts in combinatorial chemistry and its applications. The techniques of combinatorial small molecule library synthesis and high-throughput screening, pioneered in the Department of Chemistry at UC Berkeley, have enabled the rapid generation of large compound libraries for screening against biological targets. These techniques are now widely applied by pharmaceutical companies to drug discovery efforts. Combinatorial small molecule strategies to the rapid identification of natural ligands or substrates for newly identified receptors and enzymes has been proven highly effective

In order to leverage the initial capital equipment investment and be able to produce inhibitors quickly, the initial compound library to be screened will be acquired from commercial sources. Chembridge and ChemDiv sell diverse libraries of compounds prescreened for stability and bioavailability. These libraries comprise on the order of 40,000 compounds providing sufficient diversity for the identification of initial "hits" against any protein target of interest. Confirmation of hits will be accomplished by resynthesis followed by in vitro functional assay. Lead optimization will be accomplished via structural analoging (structure/activity correlation studies).

In the first 6 months, a team of experts in chemistry, biology, engineering and computer science will be assembled and charged to design the system architecture. Working together, the team must first identify (1) novel biological assays, (2) a most effective small molecule library, (3) a cost-effective, automated system for carrying out the assays, and (4) a most sensitive detection method. Most likely, some proof-of-concept experiments will need to be conducted in order to make rational decisions. Then, the team, mostly the engineers will produce a thorough, detailed system design, including a list of equipment and material needed to be purchased or developed, automation protocols and a computer infrastructure for data tracking, processing and interpretation.

3.5.3 Microbial Genomics (ORNL)

Oak Ridge National Laboratory has assembled a Microbial Genomics laboratory for the high-throughput arraying of DNA on glass slides and analysis of global cellular gene expression. The Genomics Laboratory is equipped with a robotic microarray pin spotter from Cartesian Technologies (model PixSys 5500) and the MicroGrid II biorobotic arrayer from Apogent Discoveries; the ScanArray□ 5000 Microarray Analysis System (Packard Bioscience), equipped with four lasers capable of detecting most commercially available fluorescent dyes; a 96-well microtiter plate fluorometer from Perkin Elmer (HTS7000 Bioassay Reader) for large-scale DNA quantitation; a pulse gel electrophoresis unit (CHEF) from Bio-Rad; an automatic capillary DNA sequencer (ABI 3700); a Biomek 2000 Laboratory Automated Workstation from Beckman for plasmid and PCR product purification; a Hydra HTS workstation for 96-well sample dispensing from Robbins Scientific; an Ultra30 SUN workstation for bioinformatics; and a 2-D gel system from Pharmacia. A variety of computer software (e.g., GCG, Sequencer, Gelcompare, ImaGene, QuantArray) for gel pattern analysis, DNA sequence analysis, and microarray image analysis is available. In addition to having all equipment necessary for transcript profiling in one location, the researchers who have assembled the equipment in this facility have significant expertise in transcript profiling. Thus, the researchers working in this facility will perform all transcript profiling for the proposed work.

3.6 References

1. Amann, R. I., and W. Ludwig. 1995. Microbiol. Rev. **59**:143-169.
2. Baer, A., and J. Bode. 2001. Coping with kinetic and thermodynamic barriers: RMCE, an efficient strategy for the targeted integration of transgenes. Curr. Opin. Biotechnol. **12**:473-480.
3. Bao, Y., D. P. Lies, H. Fu, and G. P. Roberts. 1991. An improved Tn7-based system for the single copy insertion of cloned genes into chromosomes of Gram-negative bacteria. Gene **109**:167-168.
4. Beliaev, A., and D. Saffarini. 1998. The *Shewanella putrefaciens mtrB* gene encodes an outer membrane protein required for Fe(III) and Mn(IV) reduction. J. Bacteriol. **180**:6292-6297.

5. Beliaev, A., D. K. Thompson, C. Brandt, C. S. Giometti, D. P. Lies, K. H. Nealson, and J. Zhou. 2002. Gene and protein expression profiles of *Shewanella oneidensis* during anaerobic growth with different electron acceptors. OMICS **6**:39-60.
6. Beliaev, A., D. K. Thompson, D. A. Fields, L. W. M., D. P. Lies, K. H. Nealson, and J. Zhou. 2002. Microarray expression profiling in Shewanella oneidensis MR-1 indicates the involvement of etrA in global gene regulation. J. Bacteriol. Submitted.
7. Blomfield, I. C., V. Vaughn, R. F. Rest, and B. I. Eisenstein. 1991. Allelic exchange in *Escherichia coli* using the *Bacillus subtilis sacB* gene and a temperature-sensitive pSC101 replicon. Mol. Microbiol. **5**:1447-1457.
8. Borrow, M. N., G. J. Litt, K. J. Shaughnessy, P. C. Mayer, and J. Conlon. 1992. The use of catalyzed reporter deposition as a means of signal amplification in a variety of formats. J. Immunol. Meth. **150**:145-149.
9. Buchholz, A., R. Takors, and C. Wandrey. 2001. Quantification of intracellular metabolites in *Escherichia coli* K12 using liquid chromatography-electrospray ionization tandem mass spectrometric techniques. Anal. Biochem. **295**:129-137.
10. Cashel, M., D. R. Gentry, V. J. Hernandez, and D. Vinella. 1996. The stringent response, p. 1458-1496. *In* F. C. Neidhardt (ed.), *Escherichia coli and Salmonella.* Cellular and Molecular Biology. ASM Press, Washington, D.C.
11. Chen, Y., E. Dougherty, and M. Bittner. 1997. Ratio-Based Decisions and the Quantitative Analysis of cDNA Micro-array Images. J. Biomed. Optics **2**:364.
12. Christensen, B., T. Christiansen, A. K. Gombert, J. Thykaer, and J. Nielsen. 2001. Simple and robust method for estimation of the split between the oxidative pentose phosphate pathway and the Embden-Meyerhof-Parnas pathway in microorganisms. Biotechnol. Bioeng. **74**:517-523.
13. Clement, B. G., L. E. Kehl, K. L. DeBord, and C. L. Kitts. 1998. J. Microbiol. Meth. **31**:135-142.
14. Coppi, M. V., C. Leang, S. J. Sandler, and D. R. Lovley. 2001. Development of a genetic system for *Geobacter sulfurreducens*. Appl. Environ. Microbiol. **67**:3180-3187.
15. Coulter, E. D., and J. D. M. Kurtz. 2001. A role for rubredoxin in oxidative stress protection in *Desulfovibrio vulgaris*: catalytic electron transfer to rebrerythrin and two-iron superoxide reductase. Arch. Biochem. Biophys. **394**:76-86.
16. Datsenko, K. A., and B. L. Wanner. 2000. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc. Natl. Acad. Sci. USA **97**:6640-6645.
17. Dauner, M., J. E. Bailey, and U. Sauer. 2001. Metabolic flux analysis with a comprehensive isotoper model in *Bacillus subtilis*. Biotechnol. Bioeng. **76**:144-156.
18. De Lorenzo, V., M. Herrero, U. Jakubzik, and K. N. Timmis. 1990. Mini-Tn5 transposon derivatives for insertion mutagenesis, promoter probing, and chromosomal insertion of cloned DNA in Gram-negative eubacteria. J. Bacteriol. **172**:6568-6572.
19. de Lorenzo, V., and K. N. Timmis. 1994. Analysis and construction of stable phenotypes in Gram-negative bacteria with Tn5- and Tn10-derived mini-transposons. Meth. Enzymol. **235**:386-405.
20. DeRisi, J. L., V. R. Iyer, and P. O. Brown. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science **278**:680-694.
21. Dilling, W., and H. Cypionka. 1990. Aerobic respiration in sulfate-reducing bacteria. FEMS Microbiol. Lett. **71**:123-128.
22. Dolla, A., B. K. J. Pohorelic, J. K. Voordouw, and G. Voordouw. 2000. Deletion of the hmc operon of *Desulfovibrio vulgaris* subsp. vulgaris Hildenborough hampers hydrogen metabolism and low-redox-potential niche establishment. Arch. Microbiol. **174**:143-151.
23. Fiehn, O. 2002. Metabolomics - the link between genotypes and phenotypes. Plant Mol. Biol. 48:155-171.
24. Fiehn, O., J. Kopka, P. Dormann, T. Altmann, R. N. Trethewey, and L. Willmitzer. 2000. Metabolite profiling for plant functional genomics. Nat. Biotechnol. **18**:1157-1161.
25. Fisher, S. L., W. Jiang, B. L. Wanner, and C. T. Walsh. 1995. Cross-talk between the histidine protein kinase VanS and the response regulator PhoB: Characterization and identification of a VanS domain that inhibits activation of PhoB. J. Biol. Chem. **270**:23143-23149.

26. Forbes, N. S., Clark, D.S., Blanch, H.W. 2000. Analysis of metabolic fluxes in mammalian cells. *In* K. Schugerl, Bellgardt, K.H. (ed.), Bioreaction Engineering: Modeling and Control. Springer-Verlag, Berlin Heidelberg.
27. Forbes, N. S., Clark, D.S., Blanch, H.W. 2001. Using isotopomer path tracing to quantify metabolic fluxes in pathway models containing reversible reactions. Biotechnol. Bioeng. **74**:196-211.
28. Fu, R., and G. Voordouw. 1997. Targeted gene-replacement mutagenesis of dcrA encoding an oxygen sensor of the sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough. Microbiology **143**:1815-1826.
29. Gavin, A.-C., M. Boesche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. Michon, C.-M. Cruciat, M. Remor, C. Hoefert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.-A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature (London) **415**:141-147.
30. Gerhardt, P., R. G. E. Murray, W. A. Wood, and N. R. Krieg. 1994. Methods for general and molecular bacteriology. American Society for Microbiology, Washington, D.C.
31. Goodlett, D. R., A. Keller, J. D. Watts, R. Newitt, E. C. Yi, S. Purvine, J. K. Eng, P. von Haller, R. Aebersold, and E. Kolker. 2001. Differential stable isotope labeling of peptides for quantitation and de novo sequence derivation. Rap. Comm. Mass Spec. **15**:1214-1221.
32. Gygi, S., B. Rist, S. Gerber, F. Turecek, M. Gelb, and R. Aebersold. 1999. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat. Biotechnol. **17**:994-999.
33. Gygi, S., Y. Rochon, B. Franza, and R. Aebersold. 1999. Correlation between protein and mRNA abundance in yeast. Mol. Cell. Biol. **19**:1720-1730.
34. Hardy, J. A., and W. A. Hamilton. 1981. The oxygen tolerance of sulfate-reducing bacteria isolated from North Sea waters. Curr. Microbiol. **6**:259-262.
35. Hegde, P., R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J. E. Hughes, E. Snesrud, N. Lee, and J. Quackenbush. 2000. A concise guide to cDNA microarray analysis. Biotechniques **29**:548-560.
36. Herrero, M., V. d. Lorenzo, and K. N. Timmis. 1990. Transposon vectors containing non-antibiotic resistance selection markers for cloning and stable chromosomal insertion of foreign genes in Gram-negative bacteria. J. Bacteriol. **172**:6557-6567.
37. Ho, Y., A. Gruhler, A. Heilbut, G. Bader, L. Moore, S. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. Willems, H. Sassi, P. Nielsen, K. Rasmussen, J. Andersen, L. Johansen, L. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. Sorensen, J. Matthiesen, R. Hendrickson, F. Gleeson, T. Pawson, M. Moran, D. Durocher, M. Mann, C. Hogue, D. Figeys, and M. Tyers. 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature **415**:180-183.
38. Huang, W., J. Petrosino, and T. Palzkill. 1998. Display of functional -lactamase inihibitory protein on the surface of M13 bacteriophage. Antimicrob. Agents Chemother. **42**:2893-2897.
39. Igo, M. M., A. J. Ninfa, J. B. Stock, and T. J. Silhavy. 1989. Phosphorylation and dephosphorylation of a bacterial transcriptional activator by a transmembrane receptor. Genes Dev. **3**:1725-1734.
40. Johnson, M. S., I. B. Zhulin, M.-E. R. Gapuzan, and B. L. Taylor. 1997. Oxygen-dependent growth of the obligate anaerobe *Desulfovibrio vulgaris* Hildenborough. J. Bacteriol. **179**:5598-5601.
41. Kasahara, M., K. Makino, M. Amemura, A. Nakata, and H. Shinagawa. 1991. Dual regulation of the ugp operon by phosphate and carbon starvation at two interspaced promoters. J. Bacteriol. **173**:549-558.
42. Keen, N. T., S. Tamaki, D. Kobayashi, and D. Trolliger. 1988. Improved broad-host-range plasmids for DNA cloning in Gram-negative bacteria. Gene **70**:191-197.
43. Khodursky, A. B., B. J. Peter, N. R. Cozzarelli, D. Botstein, P. O. Brown, and C. Yanofsky. 2000. DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in *Escherichia coli*. Proc. Natl. Acad. Sci. USA 97:12170-12175.
44. Kim, U.-J., H. Shizuya, P. J. Jong, B. Birren, and M. I. Simon. 1992. Nucl. Acids Res. **20**:1083-1085.

45. Koonin, E. V., L. Aravind, and M. Y. Galperin. 2000. A comparative-genomic view of the microbial stress response, p. 417-444. *In* R. Hengge-Aronis (ed.), Bacterial Stress Response. ASM Press, Washington, D.C.

46. Lange, H. C., M. Eman, G. van Zuijlen, D. Visser, J. C. van Dam, J. Frank, M. J. Teiseira de Mattos, and J. J. Heijnen. 2001. Improved rapid sampling for *in vivo* kinetics of intracellular metabolites in *Saccharomyces cerevisiae*. Biotechnol. Bioeng. **75**:406-415.

47. Lemos, R. S., C. M. Gomes, M. Santana, J. LeGall, A. V. Xavier, and M. Teixeira. 2001. The 'strict' anaerobe *Desulfovibrio gigas* contains a membrane-bound oxygen-reducing respiratory chain. FEBS Letters **496**:40-43.

48. Link, A. J., L. G. Hays, E. B. Carmack, and J. R. Yates. 1997. Identifying the major proteome components of *Haemophilus influenzae* type-strain NCTC 8143. Electrophoresis **18**:1314-1334.

49. Link, A. J., D. Phillips, and G. M. Church. 1997. Methods for generating precise deletions and insertions in the genome of wild-type *Escherichia coli*: application to open reading frame characterization. J. Bacteriol. **179**:6228-6237.

50. Liu, Q., M. Z. Li, D. Leibham, D. Cortez, and S. J. Elledge. 1998. The univector plasmid-fusion system, a method for rapid construction of recombinant DNA without restriction enzymes. Current Biology **8**:1300-1309.

51. Lloyd, R. G., and K. B. Low. 1996. Homologous recombination, p. 2236-2255. *In* F. C. Neidhardt (ed.), Escherichia coli and Salmonella, 2nd Ed. ASM Press, Washington, DC.

52. Macario, A. J. J., M. Lange, B. K. Ahring, and E. C. de Macario. 1999. Stress genes and proteins in the archae. Microbiol. Mol. Biol. Rev. **63**:923-967.

53. Maier, T. M., and C. R. Myers. 2001. Isolation and characterization of a *Shewanella putrefaciens* MR-1 electron transport regulator *etrA* mutant: reassessment of the role of EtrA. J. Bacteriol. **183**:4918-4926.

54. Moser, D., and K. Nealson. 1996. Growth of the facultative anaerobe *Shewanella putrefaciens* by elemental sulfur reduction. Appl. Environ. Microbiol. **62**:2100-2105.

55. Murphy, K. C. 1998. Use of bacteriophage recombination functions to promote gene replacement in *Escherichia coli*. J. Bacteriol. **180**::2063-2071.

56. Myers, C. R., and K. H. Nealson. 1990. Respiration-linked proton translocation coupled to anaerobic reduction of manganese(IV) and iron(III) in *Shewanella putrefaciens* MR-1. J. Bacteriol. **172**:6232-6238.

57. Neidhardt, F. C., and R. A. VanBogelen. 2000. Proteomic analysis of bacterial stress responses, p. 445-452. *In* R. Hengge-Aronis (ed.), Bacterial Stress Response. ASM Press, Washington, D.C.

58. Nguyen, B.-T., K. Lazzari, J. Abebe, I. Mac, J. B. Lin, A. Chang, K. L. Wydner, J. B. Lawrence, L. Scott Cram, H.-U. Weier, J. C. Weaver, and D. W. Bradley. 1995. Cytometry **21**:11-19.

59. Ninfa, A. J., E. G. Ninfa, A. N. Lupas, A. Stock, B. Magasanik, and J. Stock. 1988. Crosstalk between bacterial chemotaxis signal transduction proteins and regulators of transcription of the Ntr regulon: evidence that nitrogen assimilation and chemotaxis are controlled by a common phosphotransfer mechanism. Proc. Natl. Acad. Sci. USA **85**:5492-5496.

60. Palzkill, T., W. Huang, and G. M. Weinstock. 1998. Mapping protein-ligand interactions using whole genome phage display libraries. Gene **221**:79-83.

61. Pedersen, H., M. Carlsen, and J. Nielsen. 1999. Identification of enzymes and quantification of metabolic fluxes in the wild type and in a recombinant *Aspergillus oryzae* strain. Appl. Env. Microbiol. **65**:11-19.

62. Pohorelic, B. K. J., J. K. Voordouw, E. Lojou, A. Dolla, J. Harder, and G. Voordouw. 2002. Effects of deletion of genes encoding Fe-only hydrogenase of *Desulfovibrio vulgaris* Hildenborough on hydrogen and lactate metabolism. J. Bacteriol. **184**:679-686.

63. Pomposiello, P. J., M. H. J. Bennik, and B. Demple. 2001. Genome-wide transcriptional profiling of the *Escherichia coli* responses to superoxide stress and sodium salicylate. J. Bacteriol. **183**:3890-3902.

64. Powell, K. T., and J. C. Weaver. 1990. Biotechnology **8**:333-337.

65. Pramanik, J., and J. D. Keasling. 1998. Effect of carbon source and growth rate on biomass composition and metabolic flux predictions of a stoichiometric model. Biotechnol. Bioeng. **60**:230-238.

66. Pramanik, J., and J. D. Keasling. 1997. Stoichiometric model of *Escherichia coli* metabolism: incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. Biotechnol. Bioeng. **56**:398-421.

67. Pramanik, J., and J. D. Keasling. 1997. A stoichiometric model of *Escherichia coli* metabolism: incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. Biotechnol. Bioeng. **56**:398-421.
68. Pramanik, J., P. L. Trelstad, A. J. Schuler, D. Jenkins, and J. D. Keasling. 1998. Development and validation of a flux-based stoichiometric model for enhanced biological phosphorus removal metabolism. Water Research **33**:462-476.
69. Price, C. W., P. Fawcett, H. Ceremonie, N. Su, C. K. Murphy, and P. Youngman. 2001. Genome-wide analysis of the general stress response in *Bacillus subtilis*. Mol. Microbiol. **41**:757-774.
70. Puig, O., F. Caspary, G. Rigaut, B. Rutz, E. Bouveret, E. Bragado-Nilsson, M. Wilm, and B. Seraphin. 2001. The tandem affinity purification (TAP) method: A general procedure of protein complex purification. Methods (Orlando) **24**:218-229.
71. Robison, K., A. M. McGuire, and G. M. Church. 1998. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K12 genome. J. Mol. Biol. **284**:241-254.
72. Rose, D. J., and G. J. Opiteck. 1994. Two-dimensional gel electrophoresis/liquid chromatography for the micropreparative isolation of proteins. Anal. Chem. **66**:2529-2536.
73. Santos, H., P. Fareleira, A. V. Xavier, L. Chen, M.-Y. Liu, and J. LeGall. 1993. Aerobic metabolism of carbon reserves by the "obligate anaerobe" *Desulfovibrio gigas*. Biochem. Biophys. Res. Commun. **195**:551-557.
74. Schmidt, K., L. C. Norregaard, B. Pedersen, A. Meissner, J. O. Duus, J. O. Nielsen, and J. Villadsen. 1999. Quantification of intracellular metabolic fluxes from fractional enrichment and ^{13}C-^{13}C coupling constraints on the isotopomer distribution in labeled biomass components. Met. Eng. **1**:166-179.
75. Schwikowski, B., P. Uetz, and S. Fields. 2000. A network of protein-protein interactions in yeast. Nat. Biotechnol. **18**:1257-1261.
76. Seifert, H., and M. So. 1991. Genetic systems in pathogenic *Neisseriae*. Meth. Enzymol. **204**:342-357.
77. Setya, A., M. Murillo, and T. Leustek. 1996. Sulfate reduction in higher plants: Molecular evidence for a novel 5'-adenylylsulfate reductase. Proc. Natl. Acad. Sci. USA **93**:13383-13388.
78. Shediac, R., S. M. Ngola, D. J. Throckmorton, D. S. Anex, T. J. Shepodd, and A. K. Singh. 2001. Reverse-phase electrochromatography of amino acids and peptides using porous polymer monoliths. J. Chrom. A **925**:251-262.
79. Speel, E. J. M., A. H. N. Hopman, and P. Komminoth. 1999. J. Histochem. Cytochem. **47**:281-288.
80. Stephanopoulos, G. N., A. A. Aristidou, and J. Nielsen. 1998. Metabolic Engineering: Principles and Methodologies. Academic Press, San Diego.
81. Stock, A. M., V. L. Robinson, and P. N. Gourdreau. 2000. Two-component signal transduction. Annu. Rev. Biochem. **69**:183-215.
82. Tamayo, P., D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub. 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. USA **96**:2907-2912.
83. Thompson, D. K., A. Beliaev, C. S. Giometti, D. P. Lies, K. H. Nealson, H. Lim, I. J. Yates, J. Tiedje, and J. Zhou. 2002. Transcription and proteomic analysis of a ferric uptake regulator (Fur) mutant of *Shewanella oneidensis*: possible involvement of Fur in energy metabolism, regulation and oxidative stress. Appl. Environ. Microbiol. **68**:881-892.
84. Throckmorton, D. J., T. J. Shepodd, and A. K. Singh. 2002. Electrochromatography in microchips: reversed-phase separation of peptides and amino acids using photo-patterned rigid polymer monoliths. Anal. Chem. in press.
85. Uetz, P., L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, B. G. Y. Li, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg. 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature **403**:623-627.
86. Voordouw, G. 1993. Molecular biology of the sulfate-reducing bacteria, p. 88-130. *In* R. Singleton, Jr. (ed.), The Sulfate-Reducing Bacteria: Contemporary Perspectives. Springer-Verlag, New York, NY.

87. Voordouw, J. K., and G. Voordouw. 1998. Deletion of the rbo gene increases the osygen sensitivity of the sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough. Appl. Environ. Microbiol. **64**:2882-2887.

88. Wall, J. D., B. J. R. Giles, and M. Rousset. 1993. Characterization of a small plasmid from *Desulfovibrio desulfuricans* and its use for shuttle vector construction. J. Bacteriol. **175**:4121 4128.

89. Wall, J. D., T. Murnan, J. Argyle, R. S. English, and B. J. R. Giles. 1996. Transposon mutagenesis in *Desulfovibrio desulfuricans*: Development of a random mutagenesis tool from Tn7. Appl. Environ. Microbiol. **62**:3762 3767.

90. Warren, J. W., J. R. Walker, J. R. Roth, and E. Altman. 2000. Construction and characterization of a highly regulable expression vector, pLAC11, and its multipurpose derivatives, pLAC22 and pLAC33. Plasmid **44**:138-151.

91. Washburn, M. P., D. Wolters, and J. R. Yates. 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nature Biotechnology **19**:242-247.

92. Wittmann, C., and E. Heinzle. 2001. Application of MALDI-TOF MS to lysine-producing *Corynebacterium glutamicum*. Eur. J. Biochem. **268**:2441-2455.

93. Wittmann, C., and E. Heinzle. 1999. Mass spectrometry for metabolic flux analysis. Biotechnol. Bioeng. **62**:739-750.

94. Wright, I., J.S., and R. J. Kadner. 2001. The phosphoryl transfer domain of UhpB interacts with the response regulator UhpA. J. Bacteriol. **183**:3149-3159.

95. Yao, S., D. S. Anex, W. B. Caldwell, D. W. Arnold, K. B. Smith, and P. G. Schultz. 1999. SDS capillary gel electrophoresis of proteins in microfabricated channels. Proc. Natl. Acad. Sci. USA **96**:5372-5377.

96. Yates, J., J. Eng, and A. McCormack. 1995. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. Anal. Chem. **67**:3202-3210.

97. Yates, J., J. Eng, A. McCormack, and D. Schieltz. 1995. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Anal. Chem. **67**:1426-1436.

98. Yates, J. R., A. J. Link, D. A. Schieltz, J. K. Eng, and E. Carmack. 1999. Mining proteomes using mass spectrometry: New approaches to help define function. FASEB J. **13**:A1431.

4 COMPUTATIONAL CORE

Adam Arkin (LBNL/UCB, Core Team Leader)
Inna Dubchak (LBNL)
Terry Hazen (LBNL)
Frank Olken (LBNL)
David Stahl (University of Washington)
Judy Wall (University of Missouri)
Jizhong Zhou (ORNL)

4.1 Goals and Specific Aims

In order to interpret all the information generated by the Applied Environmental Microbiology Core (AEMC) and the Functional Genomics Core (FGC) and use it to develop new conceptual models for how microbes, in particular the three targets mentioned in Section 1, respond to stress conditions in natural environments, a sophisticated data storage and analysis environment must be designed. In this section, we describe how data flow from the FGC into a database especially designed to place experimental data and edited information into the context of regulatory networks. We then describe how these networks are to be deduced from these data and turned into statistical and causal models for pathway function. Given such models for each of the target organisms, we first describe how the dynamics and control in these networks are to be analyzed and cross-compared with one another to elucidate the conserved principles of stress response and the coordination among pathways. This comparative analysis will be further extended to deduce what might be similar and different in the stress responses of the microbial community surrounding the target organisms. This will be accomplished by examining the differences both in cis-regulatory control and the protein complement of the target organisms, as well as that which can be deduced by sequence analysis of a number of homologous genomic regions obtained by large insert cloning from the surrounding microbial population. Through comparison of these proposed different regulatory strategies to the microarrays measuring population dynamics, the different models of the regulatory dynamics can be tested and tuned. These may then be used to form the conceptual models for survival and natural attenuation/bioremediation that are the goals of the Computational Core.

The Computational Core is charged with transforming the information generated by the other two Core-groups into meaningful models of the cellular regulatory networks underlying the stress response of the key organisms. As diagramed in Figure 1.9, this involves a set of interdependent tasks. The Computational Core is charged with stewardship of the data developed by the other two cores (Figure 1.9, box 8). To serve both the experimentalists and the models, specialized pathway databases designed for efficient access and storage of “network” information will be implemented. Both experimental design and data quality will be explored in collaboration with the Functional Genomics Core. The experimental design will be optimized both for obtaining reproducible data and for producing and testing network hypotheses (box 9).

The core will also develop the tools to reverse engineer the pathway data from the perturbation response datasets generated from the Functional Genomics Core (box 9). The ideal data set for this is shown in Figure 4.1. Here, wild-type strains of an organism (wt) and its different mutants (mut) are exposed to a variety of conditions (in this case oxygen stress, carbon starvation, high metal concentrations, etc.), which may vary over time. The time-dependent response of the organism is followed by tracking the concentration (and possibly state) of as many molecular components as possible.

Ordering and clustering among these molecular responses can be used to construct hypothetical causal relations among molecular species (see the system identification Subsection below). Molecular interaction data and regulatory element prediction are further aids to developing and validating these proposed networks. Obviously, it will be too expensive to fully fill in the data cube in Figure 4.1; however, well-placed experiments and well-chosen mutants can be used to reduce the information necessary for a good prediction.

One of the hypotheses of this work is that the three bacteria, *Desulfovibrio vulgaris, Geobacter metallireducens,* and *Shewanella oneidensis*, will respond differently to perturbation in their environment and in their pathway structure (by mutation). This is almost certainly true but these differences may be unimportant artifacts of the evolutionary divergence of these organisms, or else may serve a functional role. Discriminating between these two possibilities requires dissecting first what pieces of the pathway are absolutely necessary for function. This should be conserved across the species. Second, nonconservative regulation in the three target organisms needs to be analyzed separately to determine what added control or dynamic features of the pathway are introduced by each regulatory strategy. At this point, it may prove possible to forward a hypothesis for the different

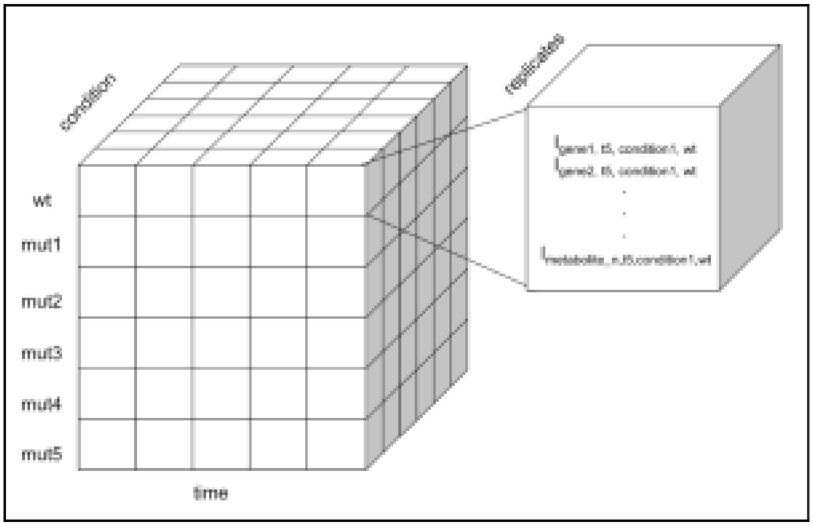

condition
replicates
wt
mut1
mut2
mut3
mut4
mut5
time
I gene1, t5, condition1, wt
I gene2, t5, condition1, wt
.
.
.
I metabolite_n,t5,condition1,wt

Figure 4.1. A diagram of the ideal perturbation/ response dataset.

functions of the different regulatory strategies (e.g., in dealing with a microaerophilic vs. anaerobic lifestyle). However, to make a robust estimation of which regulatory strategies are niche specific (or more particularly conserved among bacteria facing the same stressors), a large set of cross-comparative analyses must be done. Ideally, one would like detailed, accurate molecular models of the homologous stress response pathways in all the microorganisms in the immediate environments of our targets. However, this is too costly and time-consuming. Instead, we will clone the homologous pathways from the uncultured organisms in the same soil sample and use the resultant sequences as a basis for a comparative analysis of regulation. These large insert clones will be used both in the operon and cis-regulatory site prediction and to understand which protein and regulatory elements co-occur in species within a niche but not across niches (box 14).

The ultimate goal of this research is to create a set of detailed molecular kinetic models of stress response pathways in the key organisms, develop a practical understanding of the interplay among these pathways during different environmental conditions, and understand the comparative pathway regulation that evolved with different environmental constraints (box 10). These will constitute an unprecedented level of understanding of these generally important pathways. From these, we will develop conceptual models of the expected behavior of these populations of microbes under different stress and metal/radionuclide conditions to aid in testing hypotheses of the efficacy of natural attenuation and bioremediation strategies (boxes 15 and 16).

This section comprises a number of subprojects including Data management, Effect Detection and Parts List Determination, Pathway Deduction, and Models of Pathway Function. Each subsection contains subsections for background and significance, preliminary results, and research design and methods. The five specific aims of this the Computational Core are:

1. Create an efficient pathway data and model transport, storage and management infrastructure in concert with the other groups.
2. In direct collaboration with the Applied Environmental Microbiology Core and the Functional Genomics Core:
   a. Develop or adapt effective quality control protocols for obtaining reproducible data from the AEMC and FGC.
   b. Develop or adapt statistical data analysis for the different data types so that each observable may have two scores associated with it: (i) a binary indicator of whether or not an observed value is different from that observed in a control condition, and (ii) a real number that can be used to estimate the magnitude of the response compared both to a control condition and to other observables of the same type.

3. Develop or adapt reverse engineering tools for constructing testable hypotheses of regulatory network structure (the list of molecular players in the network and their interactions), using both comparative genomic and perturbation/response analyses.
4. Develop the high-scoring network hypotheses for each target organism into dynamical models (at different levels of abstraction) of the stress-response and reduction pathways.
5. Use the validated models to develop conceptual models for metal/radionuclide reduction in the environment

4.2 Data Management Project

4.2.1 Overview

One of the central needs for an ambitious, data-heavy project such as this is an efficient data management plan. Here we propose to develop a three stage approach to developing a database for this GTL project. In the first stage, we propose the creation of simple, custom-made flat-file databases that allow upload and retrieval by simple query of the standard data types generated by the project. Second, we will create and maintain a relational database framework to store the disparate data types in a structured format that will facilitate more complex queries and adhere to current standards of data storage and interchange made large at the NCBI and by working groups such as MGED for microarray information. Third, most of the novel work in this Core group will go into exploring the use of a class of database not yet used for biological network information—a graph database. Such databases are far better suited to queries on pathway structure and function. These goals fulfill the function of box 8 in Figure 1.9.

This latter research subproject is based on the (now commonplace) observation that many biological networks (metabolic pathways, signaling pathways, and genetic regulatory networks) can be modeled as labeled directed graphs. In addition, many queries against these network databases can themselves be represented as graphs (e.g., templates). Furthermore, many other queries can be framed as classical graph computation problems, e.g., transitive closure, connected components, graph homomorphism, graph matching, approximate graph matching, shortest paths, etc.

We propose to synthesize, adapt and extend work from the graph data management, conceptual graph processing, biological pathways databases, and graph grammars to the problem of storing and retrieving biological, chemical, and ecological network data (e.g., metabolic pathways, signaling pathways, and genetic regulatory networks). We will also develop a policy for data transfer, curation, and privacy.

4.2.2 Background and Significance

Specific aim 1 of this section calls for the creation of a unified data exchange protocol and the creation of a sophisticated database to actually hold the information. We start with a description of data exchange and move from there to databases.

Our approach to data exchange will be to create an integrated XML specification of pathways with associated data and models. In recent years, a number of biological databases have been constructed to hold and related information of various sorts. Often associated with these databases is a data exchange protocol that specifies how particular data types are to be brought into and out of these stores. These are most often XML schema. The majority of databases are sequence or structure based. The maturity of the database and XML schema in these areas reflect this. There are many different schema for transport of sequence information, including FASTA, GENBANK AGAVE (http://www.agavexml.org/, an XML specification from DoubleTwist, Inc.), and BioML (http://www.bioml.com/BIOML/). With the exception of FASTA and GENBANK, neither of which is XML, the others, although more sophisticated, have not yet found wide application. Similarly, the de facto standard for structural information is the PDB file. However, this is a rather difficult file to process, and is not XML. There are other schemes for transport of biological information becoming standards of a sort—for example, MIAMI for gene expression data (http://www.mged.org/), SBML (http://xml.coverpages.org/sbml.html) and CELLML (http://www.cellml.org/) for models and physiological information, or KeyML for taxonomic information. One important aspect of these files is that they do not easily lend themselves to preservation of a relational structure among disparate data types. What relations they do encode are not flexible enough to allow many different data types to be interrelated and annotated or to be related to a pathway context.

There are a number of excellent pathway databases however: BIND (http://www.bind.ca/), Kegg (http://www.tokyo-center.genome.ad.jp/kegg/kegg.html), EMP (http://www.empproject.com/), and Metacyc/EcoCyc (http://ecocyc.org/ecocyc/metacyc.html). Supporting databases such as BRENDA (http://www.brenda. uni-koeln.de/) contain kinetic information on reactions found in other databases. All databases differ in the amount and type of information stored, and their structure and query abilities. For example, Kegg has no kinetic or model data, whereas EMP does. Kegg is a flat-file database, which limits its query complexity, whereas BIND is a full relational schema based on NBCI's ASN.1 specification. The NCBI specification, while one of the most complete and sophisticated relational schema for biological information, is rather shallow in terms of flexibility for modeling information. EcoCyc has very detailed pathway annotations, links to expression data, and a sophisticated query language. It is, however, stored in a no-standard database with a nonstandard query language. Each of these databases differs in the availability of the underlying data schema, permissions for query, permissions for download and reformatting/reserving of data, and the underlying database technology. None of them has defined a standard format for exchange of pathway data among systems, although CellML and SBML are making strides.

Another central task is to build a novel type of database (a graph database) that supports the complex queries desirable for a modeling task. In the metabolic pathways database community, most systems have taken one of two approaches to query processing. Either they allow only very simple sorts of queries, which can be easily mapped onto underlying database systems, or they resort to ad hoc procedural programs. Some existing pathways databases only allow simple keyword (or other simple selection) queries. Peter Karp [1] has resorted to the latter technique (procedures written in LISP) for some of his work with EcoCyc and related pathways databases. As described below, we anticipate a need for complex graph theoretical queries into these pathway databases to retrieve, for example, “regulatory motifs” or find the shortest paths between two molecules. Simple queries will not be sufficient. We propose to use a special type of database called a graph database in order to support this type of query efficiently. Graph databases have been previously investigated for representing genomic data by Chris Lee and Stott Parker at UCLA.

There have been over 80 papers concerning graph data models and graph query processing. This includes work on semantic nets as described in a survey paper by Hull and King [2]. Consens and Mendelzon [3] developed Graphlog, a recursive query language on paths in graphs. Gyssens [4] described GOOD—an object-oriented graph data model. However, much of the work on XML schemas, databases, and query languages (e.g., XML query language standards) has focused on tree data models, which are often sufficient for Web documents and easier to deal with. A graph-based knowledge representation language, RDF (Resource Description Framework), has been developed by the W3C and forms the basis for a number of graph-based metadata retrieval systems.

The conceptual graph (CG) community is concerned with an approach to knowledge representation, which involves encoding first order logic (FOL) statements as graphs. Many queries against the FOL database can be answered by means of graph computations against the graph representation of the FOL database. The graph-based formalism is believed by these researchers to be easier for users to understand. CG researchers also believe that many queries can be answered more efficiently by graph computations than by logic computations (e.g., theorem proving). We plan to adopt this technology for graph homomorphism query processing, based on, e.g., the work in [5], Levinson [6] and related efforts.

Graph grammars [7, 8] are the graph analog of conventional string grammars. Graph grammar productions entail the replacement of subgraph (specified as the right side of graph grammar rule) by (usually larger) graphs (specified as the left side of a graph grammar rule). Like string grammars, graph grammars vary widely in expressiveness and computational complexity for parsing. While there have been a number of efforts to use string grammars to model the structure of DNA, RNA, and protein sequences, there has been almost no work on the use of graph grammars in biological settings. There has been some work on the use of graph grammars to model organic chemical reactions. Each of these technologies will be important for making complex pathway databases. Below we describe how each of these plays a role in our design.

4.2.3 *Preliminary Studies*

All three participants in this Core Group have experience in the building and maintaining of databases and transport schema. Inna Dubchak designed, implemented, and populated the alternative splicing database [9]and the cardiovascular comparative genomics database (http://pga.lbl.gov/cvcgd.html). Frank Olken has wide experience in database structure and query in diverse fields from spatial information to biology, and has been involved in the evaluation of biological ontologies [10-12]. Adam Arkin has developed database design and implementation for the general BioSPICE project (http://biospice.lbl.gov) and for the yeast deletion project [13].

We describe here some of this latter work to show the advantages and limitations of flat-file and relational databases for the tasks at hand. The bulk of this section will be concerned with developing the graph database technology since that requires the most explanation. We assume that most biological researchers are familiar with the very basic requirements of database management.

4.2.3.1. QuickRat Flatfile Database. The yeast deletion project recently completed a strain library for *Saccharomyces cerevisiae* that contains a large number of yeast strains, each deleted for one or both copies of a particular gene, which has been replaced by a molecular bar-code (a unique DNA sequence). Affymetrix gene chips have been made that contain the complementary sequences to each of the molecular bar codes. These chips make it possible to estimate the relative abundance of each strain in a population under different growth conditions and perturbations. The basic premise of the experiment is that if one were to drop a drug, for example, on this growing population, then those strains whose growth was especially affected by the drug would be the ones deficient in the drug target or tightly associated pathway members. To get robust measures of growth many chips had to be analyzed. The team wished for a method that allowed them to rapidly upload raw data files from the gene chip scanner, retrieve them for later use, analyze them, and query the results of the analysis. We created two different portals for this work, both built on a simple indexed flat-file architecture (QuickRat). The first of these portals allows the easy upload of chips from a web page (http://gobi.lbl.gov/~aparkin/Projects/Giaever/DTrials/Experiments/QuickRat/). The form-based entry point allows simple upload of a chip and stores the file in a predefined location, updates an index file with its location, and executes a single chip analysis script. It is then easy for the users to simply recall the chip by name and get the analyzed statistics. All the automatically generated analysis results are also stored in flat file format but are not individually queryable. When it is time to analyze a set of chips, a simple chip retrieval query is executed (by a string match to the chip index) and a form is created that allows the user to select two chips to ratio, to “quantitate,” the difference between two states (see below).

Figures 4.2 and 4.3 show the results of such an analysis that graphically summarizes the distribution of growth rates and other chip quality measures, as well as a queryable list of ratio results keyed to YPD project’s annotations and MIPS ontological functional classifications. Because these forms and queries are so simple and are based only on a simple string search, a flat-file database is efficient and effective. These were also very rapid to prototype. This trend held when more complex analyses also came on line. For example, the analyses described in Subsection 4.4 required fairly sophisticated grouping of chips into control and experimental classes, and the query interface became more complex. However, since these chips were not to be referenced to anything but themselves, there was no need to make a more complex database system.

The advantages of indexed flat-file databases are that they are relatively easy to create, and data retrieval and display are by their nature customized to the application at hand. However, when databases become large and the relationships among the data types being stored become complex, managing the flat file directory structure and the index files, doing the data updating, and creating efficient queries become difficult. Flat file databases typically lack the concurrency control and recovery facilities in relational databases. Also, if more that a few well-defined queries are required, new complex code must be developed. For a project such as this, in which the experimental researchers will be generating large amounts of data within the first year and will want facilities to easily store, backup, and query their data types, a flat-file database is an expedient short-term solution.

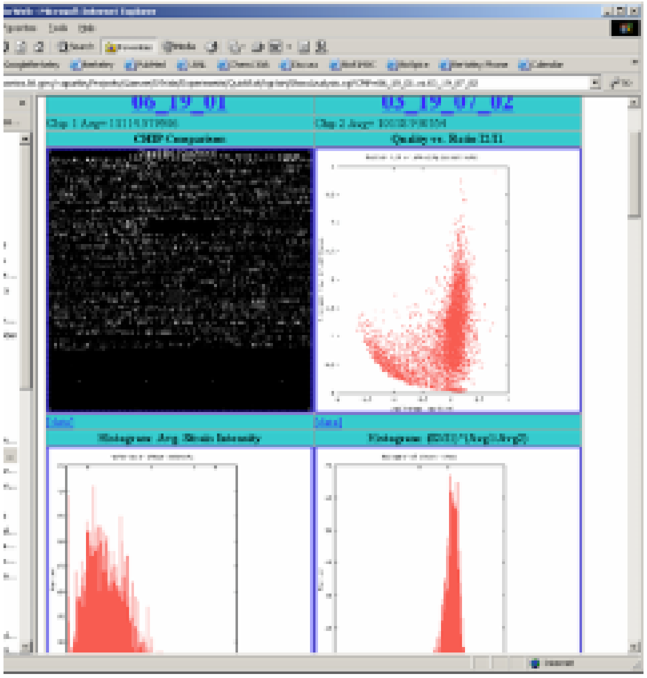
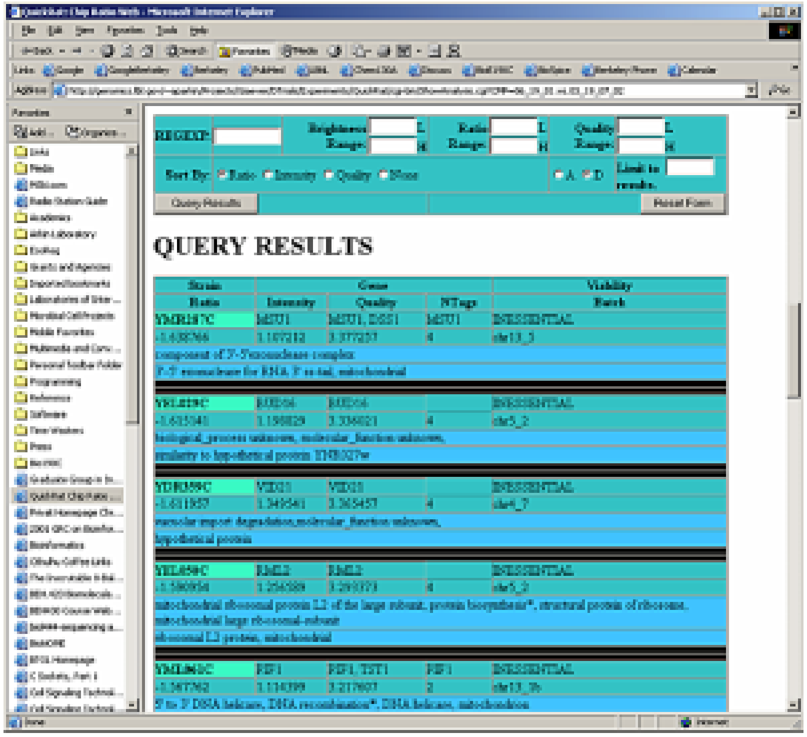


Figure 4.2. The QuickRat results display page.


QuickRat: Chip Ratio Web - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Query Results
Reset Form

REGEXP:		Brightness Range:	L H	Ratio Range:	L H	Quality Range:	L H
Sort By:	● Ratio ○ Intensity ○ Quality ○ None				○ A ● D	Limit to results:	

QUERY RESULTS

Strain	Gene			Viability
Ratio	Intensity	Quality	NTags	Batch
YMR187C	MSU1	MSU1, DSS1	MSU1	INESSENTIAL
-1.638766	1.107212	3.377257	4	chr13_3
component of 3'-5'exonuclease complex				
3'-5' exonuclease for RNA, 3' ss-tail, mitochondrial				
YKL219C	RSE56	RSE56		INESSENTIAL
-1.615341	1.190829	3.336021	4	chr5_2
biological_process unknown, molecular_function unknown,				
similarity to hypothetical protein YNR027w				
YDR359C	VID21	VID21		INESSENTIAL
-1.611957	1.349541	3.365457	4	chr4_7
vacuolar import degradation,molecular_function unknown,				
hypothetical protein				
YEL050C	RML2	RML2		INESSENTIAL
-1.590954	1.256389	3.293373	4	chr5_2
mitochondrial ribosomal protein L2 of the large subunit, protein biosynthesis*, structural protein of ribosome, mitochondrial large ribosomal subunit				
ribosomal L2 protein, mitochondrial				
YML061C	PIF1	PIF1, TST1	PIF1	INESSENTIAL
-1.587762	1.114399	3.217607	2	chr13_3b
5' to 3' DNA helicase, DNA recombination*, DNA helicase, mitochondrion				

Figure 4.3. The QuickRat results query page.

As an example of why more complex databases will be necessary, let us look at the results one of the analyses touched on in Subsection 4.4. In this experiment, the yeast population was exposed to methotrexate, a drug often used for chemotherapy. The primary molecular target of methotrexate is dihydrofolate reductase. However, the top five sensitive strains of yeast were, in decreasing order, FOL2, FOL1, YBT1, YAP6, and DFR1. The last is the target. YBT1, although labeled as a bile acid transporter in YPD (an obvious mislabel), is really a member of the ABC drug transporter class. So why are FOL2 and FOL1 more sensitive to the drug? By their gene name, it is clear that they are involved in folate metabolism. To find out how, one has to look at the pathway. A query to Kegg yields Figure 4.4. The target and the two FOL genes are circled in red. FOL1 and FOL2 are in series within a linear pathway leading to the metabolic substrate of DHFR1. Kegg does not show the allosteric regulators of these enzymes, nor does it show other connected metabolic routes between upstream metabolites and DHFR1 substrates and products. A database that could support queries that started at the five sensitive strains detected in this experiment and found all the routes among them would be very useful. (Kegg cannot do this. This pathway is only one, pre-stored route.) Certainly, if models are to be compared to growth results such as these yeast data, then efficient database access to the microarray data in pathway context is essential.

4.2.3.2. BioDB/BioSPICE Relational Database. Although the first set of databases developed will be practical flat-file databases to store the primary data coming from the AEMC and FGC, ultimately all these data will be related together and linked to the pathway information. The long-term solution is the graph database. However, implementation of a graph database is a long-term research project in itself. Thus, as a middle-term project, we propose to leverage other work on the creation of a database for a simulation and analysis tools under development in the Arkin group, BioSPICE. BioSPICE will be discussed more in Subsection 4.6. This relational database, recently ported from a MySQL to a PostGres implementation, is based on standard schema from other sources.

For a detailed discussion of the BioDB see http://biospice.lbl.gov/docs.html; we only outline BioDB briefly here. The types of data stored, and their formats, are adapted from well-known and publicly available specifications for biological information. This way we can take advantage of vast and rapidly expanding archives of existing data without investing great effort in conversion and translation. This approach also gives users the opportunity to publish their data out to such third-party databases if they choose; thus, BioSPICE avoids becoming a dead end for data, and instead benefits from its ability to use resources beyond the system itself. Currently, we inherit data and data types from several external sources:

- National Center for Biotechnology Information (NCBI): general biochemical and taxonomic data
- National Center for Genomics Research (NCGR): gene sequences
- Affymetrix/GATC Consortium: microarray experiment and chip data (this is being switched over now to conventions consistent with the MGED Group’s specifications)
- Protein Databank: protein structural descriptions
- The Biomolecular Interaction Network Database (BIND, http://www.bind.ca/)

In doing so, we have built a set of tools to automate the construction of the BioDB database from an existing third-party data specification (Schemonster, a schema translator/compiler developed for BioSPICE). Thus, new types of information can be supported automatically as new external databases are created or selected by BioSpice researchers. For more information on these databases and its schema, see http://biospice.lbl.gov/doc.html. Currently, this database holds data extracted from BIND, NCBI, data entered by our team on *Bacillus subtilis* stress response pathways, and some data on yeast protein-protein, protein-DNA interactions. Currently, all software developed for accessing these data are based on a Java Bean library or through straight JDBC calls. However, it is not yet -designed well enough to hold all the different data types to be generated by the FGC or to hold the necessary relations among the data in a structured format suitable for modeling. No Web-based browsers or query tools have been built. This will be one of the focuses of this work.

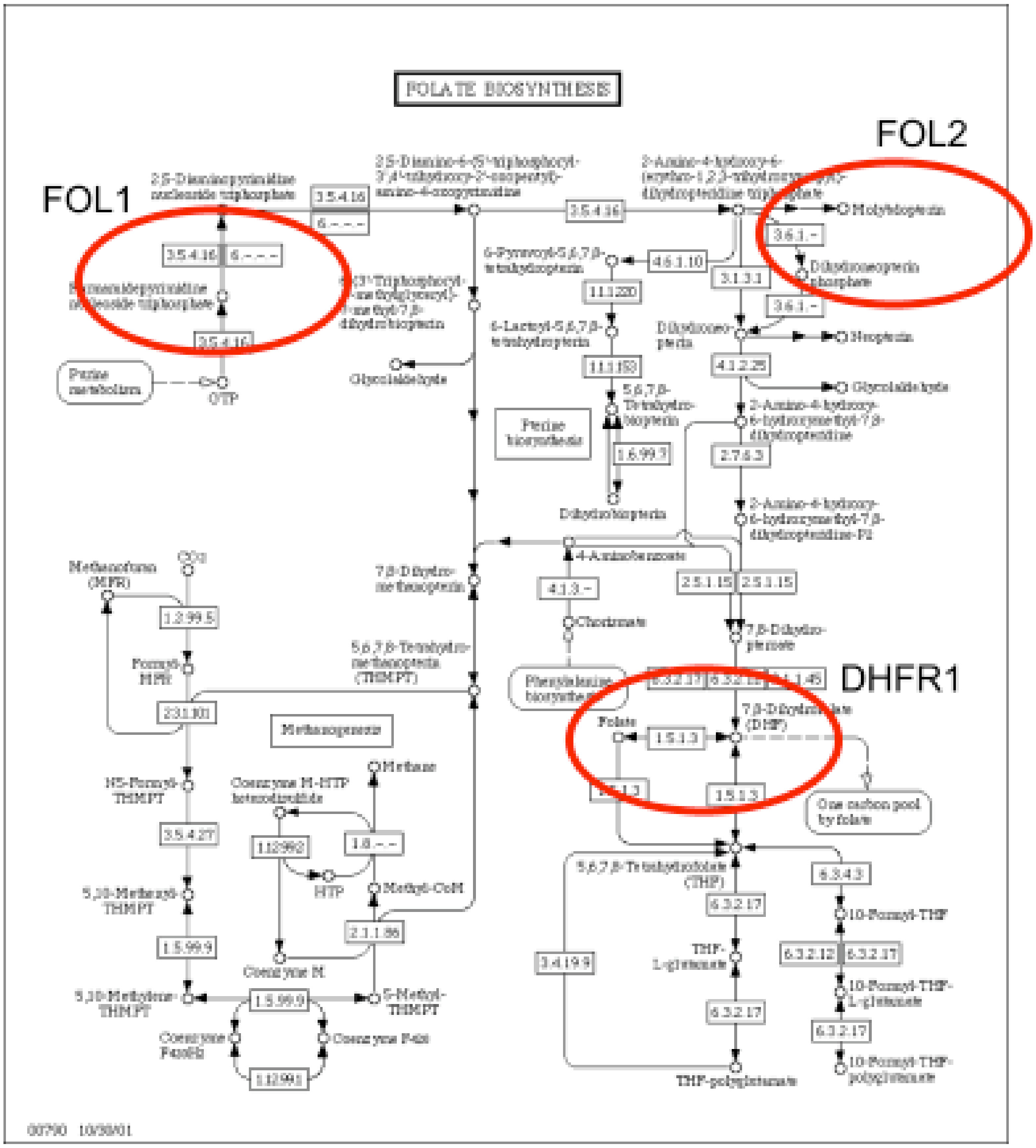


Figure 4.4. Kegg folate metabolism pathway with circled “sensitive” genes.

4.2.4 Research Design and Methods

Given the experience gained in developing the systems discussed in the preliminary results section, we foresee the need for a staged development of the final pathway database. Although our final goal is to create a database designed for efficient graphical query on biological pathways and networks, we foresee the need for early adoption of flat-file and standard relational solutions in near and midterm.

4.2.4.1. Data Exchange Protocols. The data obtained at the various sites in the FGC are either of different types or, if of the same type (such as interaction data), from different machines. In all cases, if this data are to be entered into a database of any sort, standard formats for the information need to be designed. Once these standard types have been designed, another layer of specification must be implemented that relates these data types together. We will design these transports in XML. XML is a good solution for this sort of operation because of its flexible format and protocols for syntax checking and open-source solutions for file parsing and validating. Also, many of the emerging biological data standards are XML encoded. We will have to develop standards for the following data types:

- Organism Taxonomy and Phylogeny
- Genome Data, Features, and Annotations
- Strain/Mutation/Phenotype information
- Genetic Construct Information (for laboratory management of knock-out projects)
- Biopolymer Sequence (DNA, RNA, Protein, possibly polysaccharide)
- DNA gene expression microarray
- 16S RNA microarray
- Diversa's Proteomic Data
- Metabolomic Data
- Chemical/Screen Information (Combinatorial Chemistry Facility)
- Phage-Display, Two-hybrid, Protein Cross-link Mass Spectrometry Interaction information
- In situ hybridization data
- Flow Cytometry Data
- Pathway information (in part based on interaction information)
- Data analytical methods and their outputs
- Models and model output
- A metadata framework for interrelating these data types.

As mentioned above many of these data types have existing standards. We will adopt as many of the external standards as possible before designing our own. If standards cannot be obtained externally, we will develop the XML specifications for transport consistent with the needs of the databases discussed below. In those cases where only ASN.1 data exchange formats exist (e.g., NCBI), we will mechanically convert them to XML.

In addition from the data in BioDB discussed above and the data from the FGC and AEMC, we also plan to obtain comparative information from the Shewanella Federation (which is working on *Shewanella oneidensis* MR-1) and Geobacter Consortium (working on *Geobacter sulferreducens*). Ideally, these other pathway DBs will provide bulk transfer arrangements (similar to Genbank) via XML file at ftp sites, or at least direct query APIs. If these are not available we will resort to wrapping them.

4.2.4.2. Flat-File Databases: Short Term Solutions. In the first two years, a set of Web-accessible flat-file databases of the primary data types mentioned in Subsection 4.2.4.1 will be made. These will have a similar flavor to the flat-file solutions we have already developed for Affymetrix-style yeast-deletion library microarrays. These databases will provide individual portals for uploading and performing basic data analysis on each of the separate types of data. Analyses that require more than one data type will have to download the stored files separately and analyzed them off-line.. The flat-file database exploits the hierarchical directory structure of the file system and relies on CGI scripts (in Perl) for upload, visualization, and querying. However, tools will only be used during the initial part of the project while the more sophisticated databases are developed.

4.2.4.3 Relational Databases: Adaptation and Extension of BioDB. Experience with the development of flat-file databases suggests that after the first spate of data development, the needs of the experimentalist become more sophisticated than a flat-file system can easily handle. The graph database, although probably the optimal solution, will require a longer development time-frame than the immediate needs of the FGC and modeling/data analysis researchers. Therefore, in the mid-term we will adapt BioDB to the needs of the project. This will entail a careful design phase in which all the new tables and relations for the data types and models specific to this project are constructed and their relationship to the rest of the information in the database is determined. The pathway aspects of BioDB are rapidly being organized around an object-oriented Entity-Process hierarchy, where entities can be any biological object from a cell all the way down to a region on a genome, and a process might be anything from cell locomotion to a binding reaction. Data from the different technologies are then related to which aspects of the network of entities and processes it provides evidence. For example, a spot on the 16S RNA chip might refer to the population of a particular cell and provide evidence for its growth process. However, there is another set of relations that also needs to be related to this information: the region on the genome from which the cDNA on this spot was derived, the conditions in which this experiment was run, the original raw data files, the link to the analytical method that derived the normalized spot intensity, etc. We propose to develop these schemas and populate them with project data, first by transport from the flat-file database, then, ultimately by direct upload from the FGC and AEMC sites. The relational structure of the information developed during this part of the project will also lay the framework for the graph database schema.

4.2.4.4 Graph Data Management for Biological and Chemical Networks

Overview. We propose to develop a graph-based database management system to support the storage and retrieval of cellular process network (metabolic, signaling, and genetic regulatory) data, extracellular chemical reactions, and inter-organism ecological (input-output) relationships from the AEMC and FGC. Query capabilities will include selection, projection, etc. of network paths. We will also support subgraph homomorphism queries against the network graph(s), i.e., specified as subgraph templates, with node expansion via a concept lattice. The work will encompass use cases specification, data modeling, query language specification, schema specification, query processing techniques and query optimization, graph based query interface, and interfacing to microarray data management system. The query processing prototype will use main memory query processing for the network data. We also plan to explore the use of graph grammars as an approach for representing graph queries, and potentially for representing the structure and evolution of biological network databases.

Proposed Work. Here we describe the development of the query language and query processing and optimization techniques that will be the primary focus of this sub-project. In particular, we will focus on supporting nonstandard queries such as subgraph homomorphism queries. The novel aspects of this proposal lie in the development and use of a general purpose graph data management systems for biological, chemical, and ecological pathway data, and the integration of subgraph homomorphism query facilities into a graph query language and graph query processing system.

Use Cases. We will commence by identifying a set of use cases, specifying the types of data, nature of the concept lattice (e.g., on chemical compounds) and types of queries that we expect that the project users will require. These use cases will be constructed in conjunction with all the Principal Investigators. We will also examine the publications of other metabolic pathways databases and interview some of those developers and users, e.g., BIND, MetaCyc, KEGG, Amaze, WIT, et al. A document describing the use cases will be prepared. The use cases will include a brief description of the relevant portions of the database schema, and an English language version of the queries being posed.

Examples of the use case queries include:

- Find all molecules within five reaction steps of genes overexpressed significantly as measured by a spotted cDNA microarray under specified environmental shift conditions.
- Find all molecular reaction graphs in which have the following graph structure G wherein the first two reaction nodes in the network catalyzed by kinase enzymes (kinases are a class of enzymes).
- Find all membrane diffusion processes involved in signal transduction in Prokaryotes and return the reaction networks connected to them.
- Find all signal routes starting a particular G-protein coupled receptor and ending at a particular transcription factor.

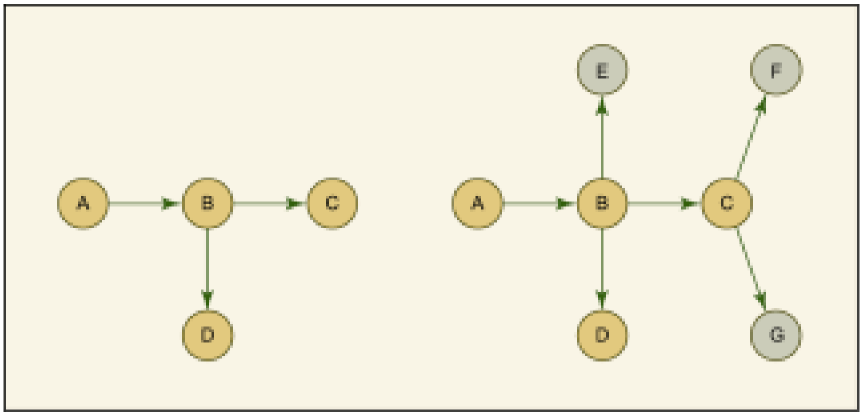


Figure 4.5. Example of subgraph isomorphism. Isomorphic subgraph in larger graph is also colored in yellow.

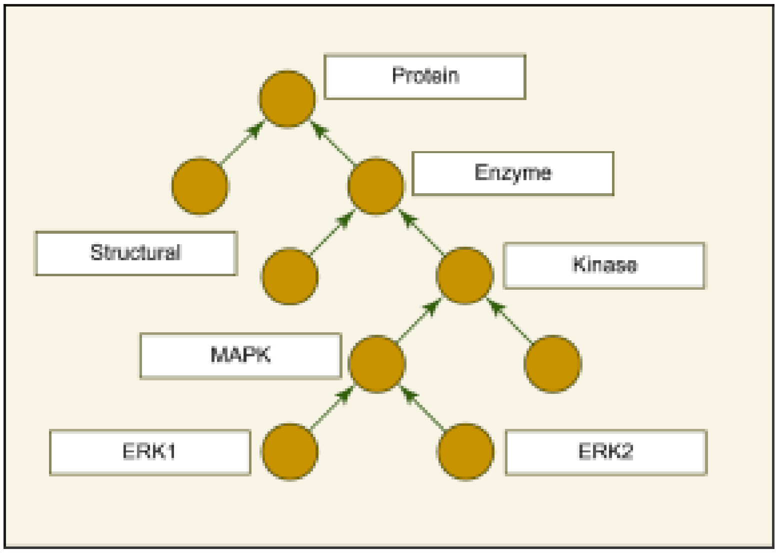


Figure 4.6. A query graph requesting all biomolecular reactions wherein chemical B is a substrate, chemical C is the product, and Enzyme catalyzes the reaction.

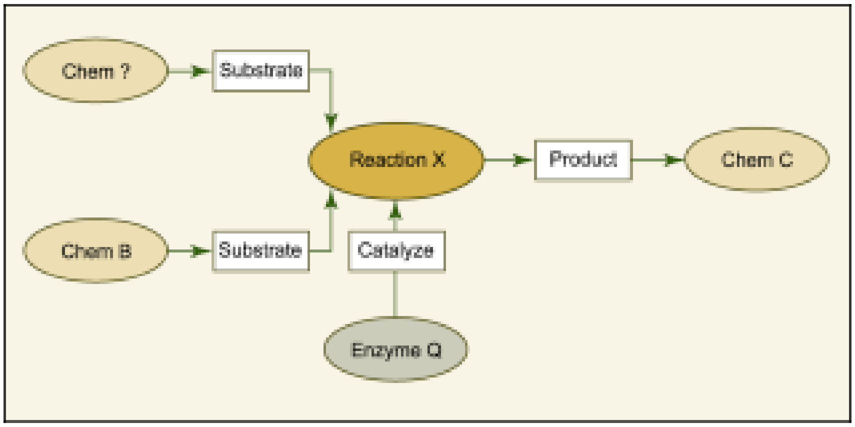


Figure 4.7. A query graph requesting all biomolecular reactions wherein chemical B is a substrate, chemical C is the product and enzyme Q catalyzes the reaction.

Queries. This project will require that we support query types not generally supported in most databases. Most notably are subgraph isomorphism (see Figure 4.5) and subgraph homomorphism queries. Such queries are typically specified as a small subgraph template (for example of a certain class of chemical reactions). The nodes of such a subgraph query template are typically specified as constraints on generic classes of reactants (e.g. abstract nouns or noun phrases). These nodes must be bound to nodes in the network database, which are specifically named reactants (or outputs). Such bindings of generic to specific nodes are mediated by means of a concept lattice, i.e., a taxonomy (Figure 4.6), which specifies subsumption (is-a) or instance (is an-instance-of) relationships among various concepts, i.e., a taxonomy.

Subgraph homomorphism queries are known to be NP-hard (i.e., the worst case computational complexity is exponential in the size of the database), since even subgraph isomorphism queries are NP-hard. There are two sources of computational difficulty: subsumption testing and node expansion (by specialization) in the concept lattice, and then subgraph isomorphism testing in the reaction graph database. Specialized encodings and query processing techniques have been developed to attack both parts of the problem.

To address subsumption (concept containment) and concept expansion in the concept lattice a number of techniques have been developed in the conceptual graph research community. Some of these techniques apply only to special classes of concept lattices. Many of them involve the construction of specialized indices or encodings of the concept lattice [5, 6]. These involve preprocessing of the concept lattice and may require recomputation of the encodings/indices when the concept lattice is changed. We believe that such "compiling" of the concept lattice will be practical for our applications, because we expect the concept lattice to be fairly stable (after initial development) except at the fringes (base) of the concept lattice. Thus we except to add individual enzymes often, but not often need to add new classes of enzymes.

Thus, some conceptual graph researchers have shown that it is sometimes (e.g., for a simple tree of concepts) possible to cleverly number the nodes of the concept lattice, and to attach an label to each concept with an interval so that subsumption relationships may be determined by simple arithmetic computations (testing the number of a node against the interval label of the putative subsuming node) rather than by exploring the concept lattice graph. This is possible if the relevant portion of the concept lattice is a simple tree. We plan to use such techniques. The alternative requires either repeatedly exploring the concept lattice graph (time consuming) or precomputing the transitive closure of the concept lattice (potentially huge). Note again that the use of such concept lattice numbering techniques requires preprocessing of the concept lattice and such numberings may require recomputing (and relabeling) under some updates to the concept lattice. As previously discussed, we believe this is manageable.

The other aspect of the subgraph homomorphism queries is the subgraph isomorphism problem on the reaction graph database (e.g., following or integrated with the concept node expansion). We note that subgraph isomorphism queries are also NP-hard in the worst case. However, chemical information systems (e.g., Chemical Abstracts Service) have effectively answered such queries for many years. A key observation is that the nodes are labeled (with element names for chemical structure graphs, and with reaction/chemical names for the reaction database) (Figure 4.7). Queries typically constrain the allowable node names. Thus, various types of indices and filtering techniques can be used to rapidly reduce the search space for the subgraph isomorphism query. Similar methods have been developed in the graph grammar research community for subgraph matching in both graph generation and graph parsing. Ohlrich et al. [14] have also reported success in subgraph isomorphism queries in CAD settings. Finally, Rudolf [15] has shown how subgraph homomorphism queries can be answered with techniques (heuristics) from constraint satisfaction problem algorithms. We will also consider these methods.

Data Model and Schema. The data model will be built on previous work in the conceptual graph community and the Protein Function and Biochemical Pathways Project at the European Bioinformatics Institute (AMAZE DB) [16, 17]. Thus, the data model will be that of labeled directed graphs. Like AMAZE, we have biochemical entities (e.g., reactants, enzymes, proteins, genes) as nodes, connected by biochemical processes (chemical reactions, catalysis, gene expression, inhibition of gene expression). To permit attribution of the biochemical processes they will also be represented not simply as directed edges, but rather as a pair of directed edges with an intervening vertex which would carry the name of the relationship (Figure 4.7). This sort of schema is widely used in the conceptual graph community and in the AMAZE project.

Query Processing and Optimization. Work on query processing and optimization techniques for various types of queries against biological network databases is the core of this research. The success of this approach to biological network databases hinges on our success at deploying effective query processing techniques for the complex queries we anticipate. As described above we plan to adopt techniques from the conceptual graph (CG) and graph grammar (GG) communities for efficiently processing of subgraph homomorphism queries. However, we must also be prepared to support more conventional select, project, join queries and a number of other unusual graph queries. For example, there have been interesting questions posed as to the diameter of metabolic pathway graphs. The diameter of a graph is the length of the longest shortest path between any two nodes (directed path in the case of directed graph).

We have also noted interest in computing k-shortest paths between two nodes in a pathways graph. Other possible queries involve more complex, but still incomplete specification of subgraphs—with the query system to fill in the missing portions of the query graph from the network database. Apropos this topic there has been work on approximate graph matching [see [18] and http://www.cs.nyu.edu/cs/faculty/shasha/papers/agm.html.]

To foster simplicity of implementation and performance, our prototype will rely on main memory query processing techniques. Persistent storage of the database will use either a relational DBMS or an OODBMS. We will commence by simply importing the network DB into main memory and then processing the queries there. We may push some selections and projections into the underlying DBMS.

The underlying DBMS will reside on the server, and the user interface will run on the user's client machine (for interactivity and scalability). However, the graph query processor could run in either the server or client machine. The attraction of running the graph query processor on server is that for small numbers of users this will provide better performance since less data will transfer over the network. By running the DBMS and graph query processing on the same machine, data transfers can be quickly accomplished in memory. Running the graph query processing in the client machines, however, offers better scalability since there is likely to be a small number of DBMS servers, and potentially many more user client machines. If needed we could resort to a three-tier architecture, with the graph query processing running in the middle tier (i.e., a separate machine from the DBMS server(s), usually co-located on a high bandwidth local area network). Initially, we will not support large numbers of users for this project. Hence, we will run the graph query processing on the DBMS server machine, directly invoking the DBMS from the graph query processor.

The user interface, however, will run on the users' client machines. This will help scalability some, and offer better interactivity at the UI. We will connect the clients to the server via the XML-based SOAP protocol (see http://www.w3.org/TR/SOAP). We would then encode (serialize) the graphs sent between server and client in XML. XML and SOAP tooling is increasingly available. The client is responsible for drawing the graphs and rendering them.

Query Language and API. Most users will likely interact with our pathway database system via either forms-based or graph-based query interfaces (discussed below). Nonetheless, providing a text-based query language is important. Such a query language will likely be the means by which various query interfaces or other applications, e.g., simulation programs, interact with the pathways database. A common alternative to provision of a full query language is to define a more limited library of common parameterized queries or a navigational interface. Neither approach is an adequate substitute for a full query language.

As noted above, there has been work in the DBMS community on query languages for graph databases, e.g., Graphlog [3]. However, we do not believe that existing graph query languages are sufficiently expressive for this project's applications. One of our first priorities will be development of more expressive graph query languages. In particular, we plan a query language capable of posing subgraph homomorphism queries. We also plan to support k-shortest path and distance constrained queries. We will provide a means to specify queries that span both the biopathways database and the FGC primary database (likely to be a conventional relational DBMS).

We have not yet settled on such a graph query language. However, it is likely that it will include some sort of recursive path expression mechanism, similar to that proposed in Consens and Mendelzon [3] or more recently adopted in the XPath query facility for XML (see http://www.w3.org/TR/xpath20) that is part of the more extensive XQuery specification (http://www.w3.org/TR/xquery/). Although our graph data models are somewhat more general than the XML data models, and our target queries, e.g., subgraph homomorphism, are also much more complex, we have been favorably impressed by the use of functional query language ideas in XQuery and the role of functional programming languages in implementations. We therefore plan to construct a functional query language and plan an implementation (in main memory) via a functional programming language (ML or Haskell). Note that YATL (a query language for XML) and Kliesli CPL (a query language for database integration used in bioinformatics settings) have both adopted this approach. YATL was implemented with CAML (a dialect of ML), and Kleisli was implemented in ML.

There are several advantages to functional query and programming languages. Functional query languages are readily amenable to integration with functional programming languages, e.g., incorporating computations within queries. Second, functional languages have proven readily amenable to parallelization and a variety of optimizations. Third, functional programming languages often include direct support for pattern matching. This is useful for query optimization and (potentially) for graph searching.

Should both Haskell and CAML prove impractical (i.e., in terms of performance) we would fall back onto an imperative programming language such as C++ or Java.

Human Query Interface. Work by a number of investigators in the DBMS, hypertext, and pathways communities suggests that a query interface based on graph templates will be attractive to users. In a graph-based interface the user sketches (with a graph editor) a template (skeletal) graph to represent the query. Constraints on node labels are then added either by means of pop-forms attached to the nodes, or by linking query template graph nodes to nodes in the concept lattices (e.g., individual taxa in a taxonomy of chemical compounds). The system would convert this to a textual query and ship it to the graph manager, which would then search the graph database (and concept lattice) to satisfy the query. The result would be passed (as an XML file) to graph drawing program, which would pass its results (as an XML SVG file) to a Web browser for viewing. There are many publicly available graph editors that might be adapted to this purpose. Another option is to link it to the BioSPICE pathway builder/display technology described below.

Graph Grammars. Graph grammars are the analog of string grammars for graphs. They [7, 8] have been researched for applications in protocol representation, concurrency control, visual language design, and software engineering. We believe that graph grammars will prove to be attractive for describing graph query languages. (They have been used to characterize graphical graph query languages.) We also believe that they will prove useful in specifying patterns in biological networks and possibly in modeling the evolution of biological networks. Graph grammars also offer a potential vehicle for hierarchical description of biological networks. We plan to explore these issues together with other project researchers.

Computing Platforms. Development will be done initially with a Unix-like environment. Where possible, we will use completely portable versions of the programming languages. Our initial development machines will each be one of Linux, FreeBSD, IRIX, or Solaris. Interface to the database will be through direct use of a provided programmers API, or through Web-based forms or Java applications.

4.2.4.5. Data Entry, Access, and Revision Policy. One of the central problems facing any database development project is how and when to incorporate new data, how to quality control this information, who has authority to enter, edit, and remove information from the database, and to specify how and when information is made public. These problems all fall under the general rubric of curation. Twice a year, the steering committee will meet to discuss the status of the database, obtain reports from the FGC heads on the quality of the information and utility of the software, and provide guidance to the database team for further development of the framework.

For the flat-file databases, data entry forms for each type of information will be limited to the laboratory that developed this information. Every user of the system will have his or her own account and all data have an attached UserID of user who uploaded it. A user/group/world protocol for publishing data will be in place. Write authority can only be granted to user and group. Read authority can be granted to anyone, and access control lists will be implemented to allow individual access to be granted.

Once data are uploaded, editing will be allowed in the following ways: (1) read and write permissions may be modified; (2) data may be deprecated only by the original developer of the information; (3) quality ratings may be given to the data by anyone; (4) revisions may be made that do not overwrite the original information but are the first displayed upon query; and (5) text annotations may be added by anyone but are displayed separately from the user curated information.

A similar protocol will be in place for the other two database implementations, however here, because of the information that must be entered and evaluated on the relationship among the data, area curators will be appointed who are responsible for periodically examining all the new data entered since last check and ensuring the quality of that information. Judy Wall will be the *Desulfovibrio* curator, Jizhong Zhu will be the *Shewanella oneidensis* curator, and Terry Hazen will be the *Geobacter metallireducens* curator. Each of the lab heads for the data generation labs in FGC will also be responsible for examining and rating new data entered into the database and reporting to the steering committee. Similarly, Arkin and Dubchak will be responsible for quality controlling the information produced by the bioinformatics and modeling tools. To aid in curation, every data type entered into the database will have a field indicating the responsible reviewer. Monthly, reports on the new and changed information will be sent to the reviewers. In addition, the Technical Advisory will provide periodic guidance on how to improve the facility.

Data will be published in the database on or before publication of related papers. Genomic sequences will be deposited in GenBank where appropriate. All data will be available by query at our web site (e.g., via forms), and a query API (e.g., SQL or our graph query language) will be provided. Data exports will be via XML as described.

4.3 Effect Detection and Parts List

4.3.1 Overview

One of the primary goals of any functional genomics analysis is to detect which molecular species are involved in the execution of a particular cellular behavior or response. For example, one might want to use a microarray experiment to identify the target for a particular drug. This requires that each atom of data (spot on a microarray, channel on a flow cytometer, peak in mass-spectrum) has a statistical model that is capable of telling how much an observed value differs from some standard and if this difference is significant. When trying to construct network hypotheses, these data provide estimates for what molecular players may be important to include in the model. In this subproject, we describe approaches for enumerating the different molecular parts that play important roles in the stress response pathways. In addition, the general principles of the data quality control protocols are discussed.

4.3.2 Background and Significance

For the data developed by the AEMC and FGC to be useful for the deduction and modeling of stress response pathways, they must be carefully quality controlled and scored against standard conditions. Without the data in hand, it is often difficult to propose optimal methods for analysis of this information. In lieu of this, we will describe our previous experience with analysis of data such as that proposed in the FGC section. We will be taking two complementary approaches to compiling a list of the molecules involved in stress response. The first is to analyze the perturbation-response information developed by the FGC to identify those molecules whose activity is significantly changed during the exposure to different stressors. The second is a comparative analysis of intergenic sequence proximal to co-expressed genes or of regions homologous within a number of different microbes to discover cis-regulatory elements. Also in this task are operon reconstructions that catch linked proteins not discovered during the perturbation response analysis.

The first task is a canonical in biological database analysis and has a long history. There are a plethora of papers on, for example, analysis of microarray data to find significantly changed transcripts (e.g., [19-23]. Examples of classifying the response of gene expression to perturbation via clustering can be found in Figures 3.2 and 3.3and the cited references therein. As Finkelstei et al. [19] point out, no normalization method addresses all the different chip artifacts that arise, such as spatial variation, blotching, pin-set differences, hybridization difference, etc. Thus, any approach to analyzing gene expression chips should try to employ different normalization schemes before the results are used for further analysis. Each experimental technology has different factors that must be considered by calling a particular feature as significant. One of the major challenges of this work is to ensure these factors are enumerated and dealt with for each data type. Data types, such as two hybrid experiments that measure protein-protein interactions, need especially careful analysis. For example, in two recent comprehensive yeast two hybrid screens

[24, 25], there very little overlap was found between the two sets. There are many arguments as to how each significant interaction was called or rejected or if there was selection of targets (e.g., [26]. In this case, using multiple modalities for interaction (in our case, phage display, two-hybrid and protein cross-link mass spectrometry) may aid in adding confidence to the data.

The second task is a relatively recent innovation. With the exception of the use of homology to assign function to proteins, comparative genomics for operon prediction [27-29], regulatory element detection [30-35], and molecular interaction [36] are only recently in common usage. The use of the phylogenetic profile in which proteins are clustered by which organisms orthologs are present [37], and other phylogenetic annotation tools [38] have become possible as more genomes have been sequenced. All these methods have their own false positive and false negative rates but are excellent complements to molecular profiling technologies. The combination of the two approaches should be a powerful method to create testable network hypotheses upon which models can be built.

4.3.3 Preliminary Studies

4.3.3.1. Haploinsufficiency Profiling in *Saccharomyces cerevisiae.* The Arkin laboratory has been involved in a number of molecular profiling projects. One study, analogous to the proposed population tracking studies, is with the yeast deletion project described briefly in Subsection 4.2.3.1. Our approach to quantifying the growth defect for each strain in the population is as follows. Each strain of yeast has two molecular bar-codes in place of the deleted gene. These are 25 nucleotide sequences chosen to be orthogonal to one another (minimal predicted cross-hybridization), and have similar physical properties like melting temperature and no secondary structure. One tag is called the up-tag and the other the down-tag based upon its placement relative to a resistance gene in the knock-in cassette. The gene chip bears the complement to both the Watson and Crick strands of these bar codes as well as tags each containing a single base mismatch in the center of the oligomer. These latter tags are supposed to serve as cross-hybridization controls. Thus, every strain of yeast has eight tags

A pool of yeast (containing equal amounts of all the deletion strains) is inoculated into YPD and sampled every 0.5 generation times. The genomic regions containing the bar-codes are PCRed out of the genomic prep of these samples using common PCR tags on either side of the insert. The resultant (labeled) DNA is hybridized to the gene chip, washed, and scanned to find fluorescent intensities for each tag. Generally, the higher the intensity, the more of that tag was in the original population and the better that strain competed for resources in the perturbed environment. The question is how to compare the intensities of different tags to obtain the relative amounts of each strain in the population.. It took six months of quality control on the experimental protocol to achieve a chip to chip correlation of greater that 0.99. The readout of most tags was highly reproducible; a few tags proved unreliable.

After a number of different standard analyses of the chips were tried (such as simple ratio tests) and found susceptible to noise, a final method was finally adopted wherein a standard condition (unperturbed growth in YPD) was replicated 10–20 times. From these runs, the statistical distribution of tag intensity over the replicates was estimating by fitting to a log-normal distribution. Intensities for each tag, measured under perturbed conditions are scored as the probability of being drawn from that control distribution. The score is a log-likelihood. Figures 4.8 and 4.9 shows the results for two different tag sets in which the condition did not affect the strain shown. Figure 4.10 shows a condition in which the strain shown in Figure 4.8 is especially sensitive.

The tags for the YHR007C are all relatively high intensity, and they have small variances. All the mismatch tags are of lower intensity than the match tags. These are hallmarks for a reliable tag. Small changes in the observed intensities are significant. On the other hand, the tag distributions for YPR143W have high variances. It would take a large change in observed intensity (compared to the control mean) to score that change as significant (that's why ratios don't work). Thus each tag has its own statistical model against which an observed intensity is scored.

The actual score given to a particular strain is the mean of the log of the probability of observing the intensity under each tag distribution. Figure 4.11 shows the results for a small subset of the conditions in our database. Each column in the plot is a different experiment (a drug, at a given concentration, after some number of generations of

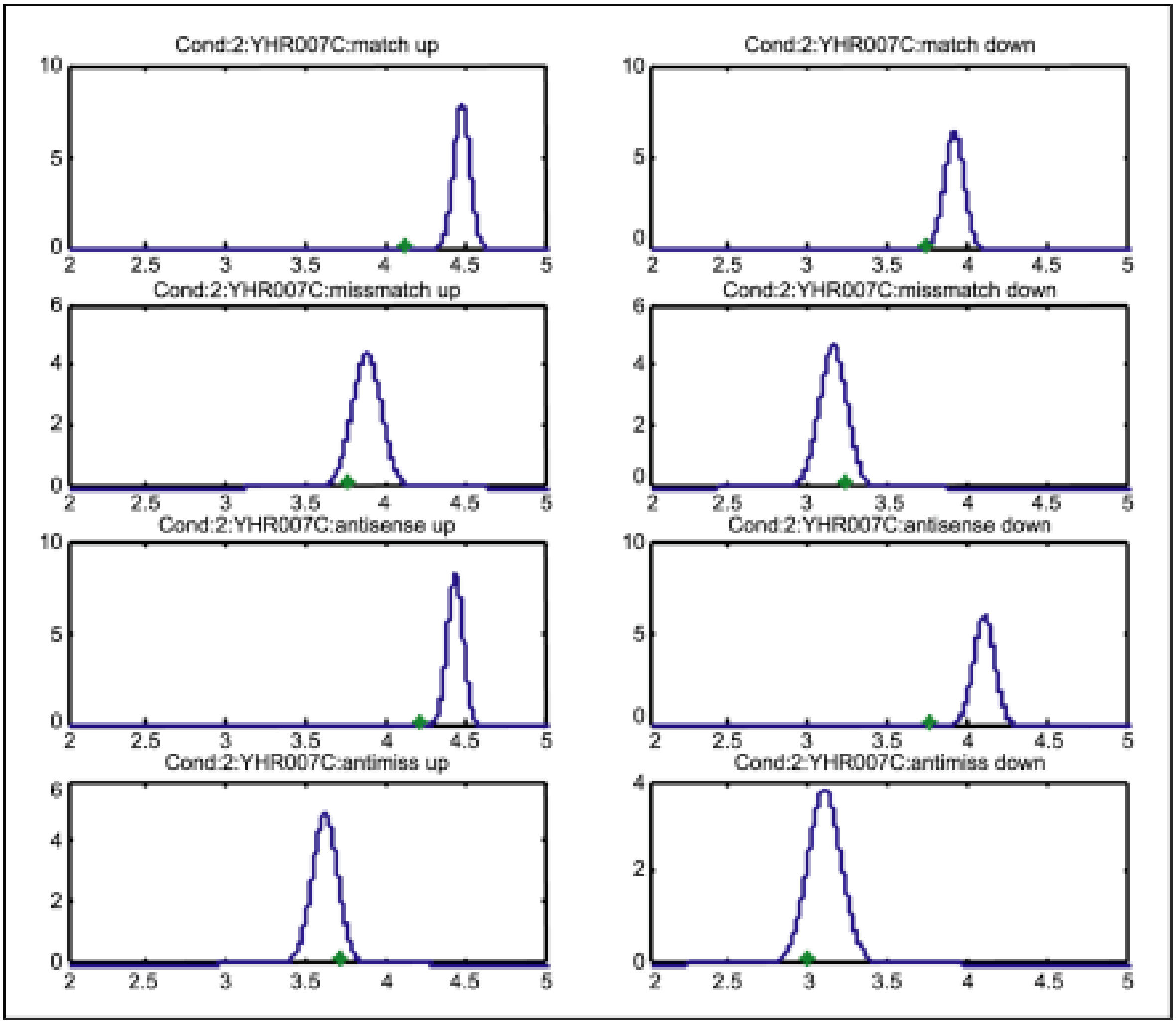


Figure 4.8. The tag intensity distributions for each of the 8 tags for strain YHR007C. The x-axis on these plots in the log of raw intensity. The y-axis is probability density. The red dot is the tag intensities for a particular experiment in which this strain was not strongly affected.

growth). Each data point in a column is a single strain measured by the chip, and its position on the y-axis is this mean log-probability. Notice that unlike gene expression data, only a few strains show a major deviation from average behavior under any given condition. The circled column is for the methotrexate results discussed in Subsection 4.2.2.1 above. Columns in which the data are compressed to the bottom of the y-axis are experiment where the drug dosage was too low to have an affect. In all cases, the drug target turned up in the top ten.

Interestingly, when various measures of dependence are performed to compare this sensitivity data to gene expression, little relationship is found. That is, the genes that change expression by large amounts are not necessarily the ones most responsible for the cell's survival in a given condition. Notable exceptions to this are conditions that result in large transcriptional changes such as shift from glucose to galactose utilization. However, almost no drug condition, or conditions of high salt, high alkali show any form of correlation. This underscores the need to measure as many molecular species as we can to understand the system response. This example is meant to demonstrate the efforts necessary to develop a robust data model sufficient for scoring molecular profiling data and cross-comparing the results.

4.3.3.2. Comparative Analysis. Adam Arkin's lab has recently installed a retooled version of the phylogenetic profiling and Rosetta Stone method pipelines on his computers. These have been applied to exploring the

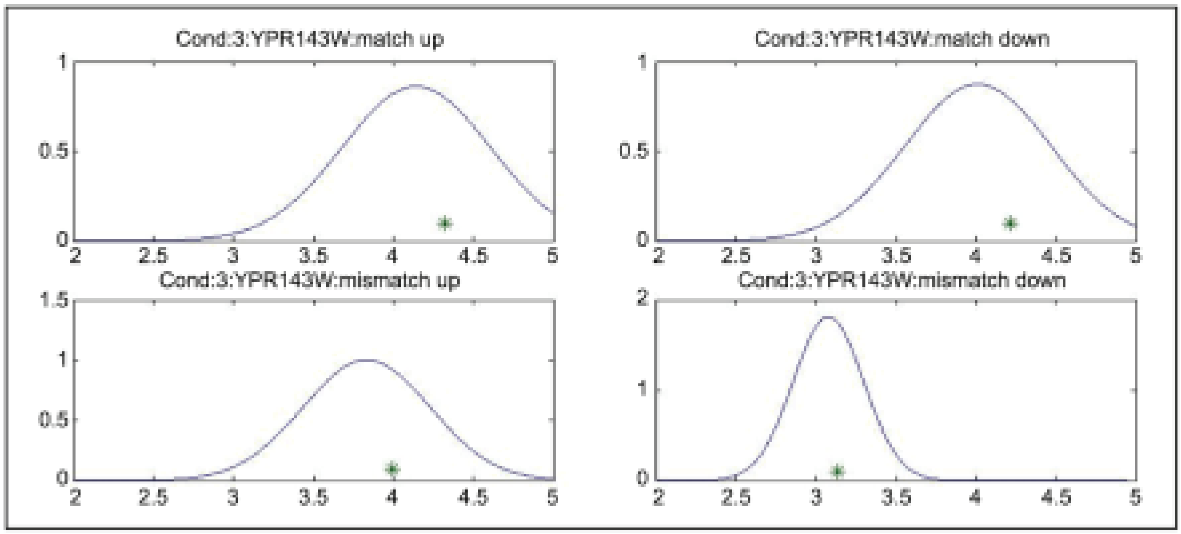


Figure 4.9. The tag intensity distributions for four of the tags for strain YPR143W.

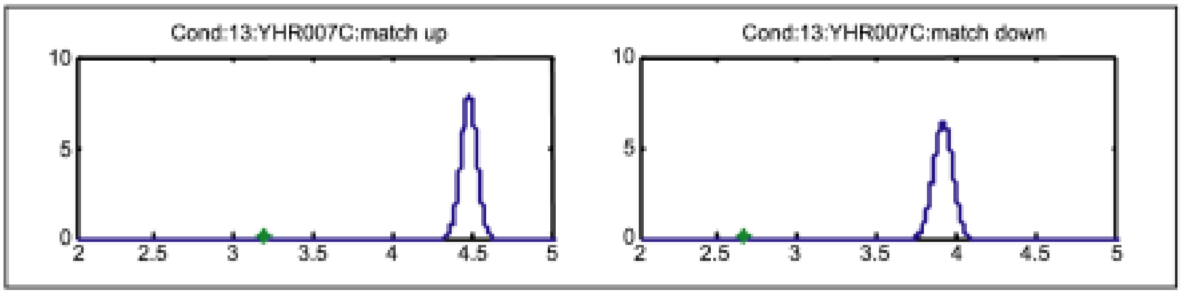


Figure 4.10. An experiment in which YHR007C is very sensitive. Notice that the red dot is well left of the bell of the distributions in each of the tags (only two tags shown).

connection between the minD and minE system in *B. subtilis* and the newly discovered bacterial "cytoskeleton." However, these results are too preliminary for even this section. They are mentioned to note that these tools are easy to install and run.

Inna Dubchak's laboratory has long experience in developing and implementing comparative genomics tools. Recently she has been developing cross-species genomic sequence comparison tools. For analyses like Dr. Dubchak's, complete genomic sequences must be accurately aligned, representing a very different problem from that solved by local alignment algorithms. The comparison of two orthologous genomic intervals requires global alignments of sequences.

Alignments of whole genomes have already been implemented for complete genomic sequences of various bacterial species [39, 40]. The whole genome alignment of several bacterial species has been feasible due to the small genomic size (up to 4 Mb) of these organisms, and the availability of strains that are very closely related to each other and share significant sequence homology [41].

One of the few as well as the most extensively used approaches for performing global alignments of large genomic regions for the purpose of cross species comparisons is PipMaker, which computes alignments in similar regions in two DNA sequences [42]. PipMaker and its graphical display are primarily targeted at identification of

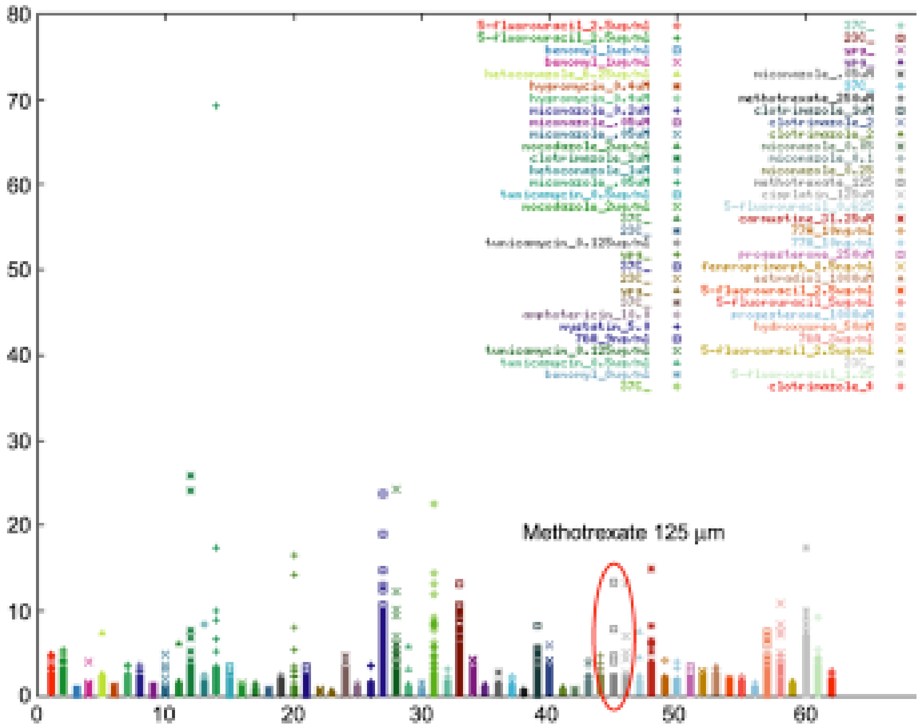


Figure 4.11. Sixty-two experiments, one column for each.

open reading frames (i.e., intron exon boundaries) and not the identification of conserved noncoding elements. In response to the needs of comparing genomic sequences from multiple organisms coupled with an interest in identifying conserved coding as well as conserved noncoding intervals, we have developed an integrated package that includes global alignment, analysis, and graphical display tools called VISTA (Visualization Tool for Alignment). Both the alignment system and the graphical display tools were built for the purpose of identifying conserved noncoding sequences in orthologous genomic intervals from two species. These informatics tools and their application to cross-species genomic sequence analysis are described in the preliminary results.

The problem of speed for whole genome alignment has been addressed by two different approaches: anchor based alignment, e.g., MUMmer [41], GLASS [43] and AVID [44], and seeding of alignments, e.g., the BLAST [45] or BLAT [46] suite of tools. The fast extraction of exact matches between sequences seems to be a fundamental building block of fast alignment algorithms, and the above approaches all rely on hash based, or suffix tree methods for this purpose. Thus, the basic "trick" of searching for exact matches seems to be the method of choice for speeding up alignment problems. A more serious difficulty that is not easy to overcome using basic heuristics is that of rearrangement, repetition and duplication in whole genomes. These features of genomes become important and difficult to address in the context of whole genome comparison.

The seeding methods overcome these problems by being inherently "local alignment" algorithms. Thus, all similarities are found, not only contiguous ones. The problem with local alignment is that the added robustness of local alignment comes at the price of less specific matches. For example, if a gene has paralogs in one organism, all these paralogs will be aligned to one copy of the gene in the second organism. On the other hand, anchor based alignments which avoid the false positive problem described above, suffer in their inability to correctly align sequences which have been subjected to rearrangements or inversions, and extra pre or post processing is necessary to identify and compensate for such occurrences. In any case, it remains an open problem and a significant challenge to build a single program that will correctly identify duplications and rearrangements, and integrate this process seamlessly into the alignment.

Nonetheless, we have developed and extensively tested an integrated system for global alignment and visualization, specifically tailored for the needs of comparative genomic analysis. We have developed and tested a number of tools that allow us to overcome many of the problems mentioned above. We have built an alignment system developed specifically for the purpose of annotation and biological discovery using orthologous genomic sequence from two species. We based our method on the AVID program [44], the program superior to its predecessor GLASS Global Alignment SyStem [43, 47] developed by one of our collaborators Lior Pachter.

The AVID alignment program works by iteratively aligning large genomic sequences at increasingly finer resolution, somewhat analogously to the MUMmer program, but with much more stringent criteria for the fixing of positions in the alignment. The aligner starts by finding large exact matching sequences between the two input sequences. These exactly matching sequences are aligned to eliminate spurious random matches. In these alignments, the matching segments that are favored are those with flanking regions that are very similar between the two input sequences. In other words, the first step of the algorithm tries to find long exactly matching sequences that are also conserved beyond the identical region, and fixes those in the global alignment. Having fixed the alignment for highly conserved regions, the aligner proceeds to align at progressively finer resolution, each time repeating the procedure outlined above, and eventually producing alignment at the base pair level. AVID works both with finished and draft sequences. In the case of draft sequences the program can automatically define the order and orientation of draft contigs, and globally align them. The AVID algorithm has been extensively tested as a module of the VISTA tool and proved to be robust on various pairs of species from human/primate to human/mouse.

To complement our global alignment program, we have also developed a graphical tool for analyzing alignments, called VISTA (Visualization Tool for Alignment), which distinguishes itself from previous attempts in a number of respects. The visual output is clean and simple, allowing the user to easily identify conserved regions. Similarity scores are displayed for the entire sequence, thus allowing for the identification of shorter conserved regions, or regions with some gaps. Figure 4.12 shows VISTA alignment results for human/mouse pairs. The highly conserved noncoding elements A and B are easily identified from this analysis.

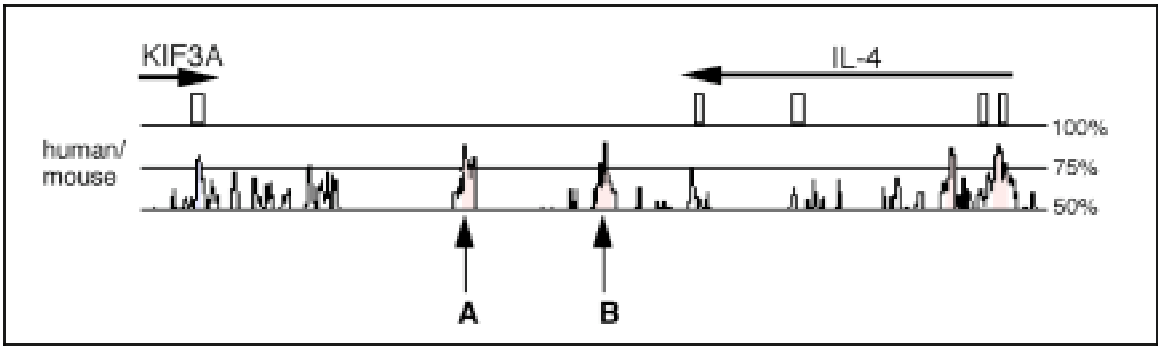


Figure 4.12. VISTA plot demonstrating peaks of similarity in the aligned Human/Mouse orthologous sequences. The horizontal axis represents the human genomic sequence while the vertical axis indicates the percent of identical nucleotides between human and mouse across the alignments (100 nucleotide window, moved by 40 nucleotide increments). The locations of coding exons, 5' and 3' UTRs are shown as rectangles above the profile. Horizontal arrows indicate the direction of transcription for each gene. The two vertical arrows, A and B indicate the peaks representing highly conserved noncoding regions. Interestingly, the exons of IL4 are significantly less conserved than the two noncoding elements depicted by the vertical arrows A and B. The plot is created by a Java program which outputs Acrobat PDF files. The X-axis of the plot represents the human base; the Y-axis represents the percent identity at that base in the alignment. Genes are marked above each plot, with the directionality indicated by the arrow. Exons are marked with dark blue rectangles above the plot and UTRs are marked with light blue. Conserved, noncoding regions are marked with red rectangles in the plot.

There are various modifications of VISTA for solving particular biological problems: cVISTA (complementary VISTA) is used for looking at differences between recently evolved species such as comparing mice to rats or humans to chimpanzees; multiVISTA allows to visualize several related alignments on the same scale. When orthologous sequences of three species are available, we can also apply a statistical method for calculating cutoffs to define noncoding sequences that are conserved because of functional constraints [32].

VISTA is implemented as a Web server and a stand-alone package. The total number of queries (pairs of sequences for comparison) submitted to the VISTA Web server exceeded 7000. Average monthly use in last three month is around 800 queries. Both AVID and VISTA are licensed through the LBNL office of technology transfer. More than 400 of the VISTA software packages have been distributed in 30 countries.

4.3.4 Research Design and Methods

4.3.4.1. Analysis of Molecular Profiling Data. Our approach to analysis of the molecular profiling data described by the FGC is to approach it like we did the yeast haploinsufficiency data. This means, the data analysts will walk through the experimental process with each data generation laboratory and develop a rough analysis of variance to identify and correct major sources of irreproducibility. These will be experimentally dealt with by the FGC lab leaders. Once reproducible results have been obtained, a scoring system like that derived for the yeast system above is possible. However, because of the complexity of the different data types we will simply have to assert that such a quality control and scoring method will be found. Thus, what we propose in this section is to start by helping design the experimental pipeline described in the FGC, aiding in the data transport and storage tasks above, after which we will examine the data to produce estimates of the number of replicates and sample sizes necessary in creating such a score. Our prior work at analyzing data of similar types gives us confidence that good measures will be found.

4.3.4.2 Comparative Analyses. Comparative genomics analyses can reveal a great deal about the important properties of regulatory networks. For example, analysis of the operon structure of multiple organisms can lead to a

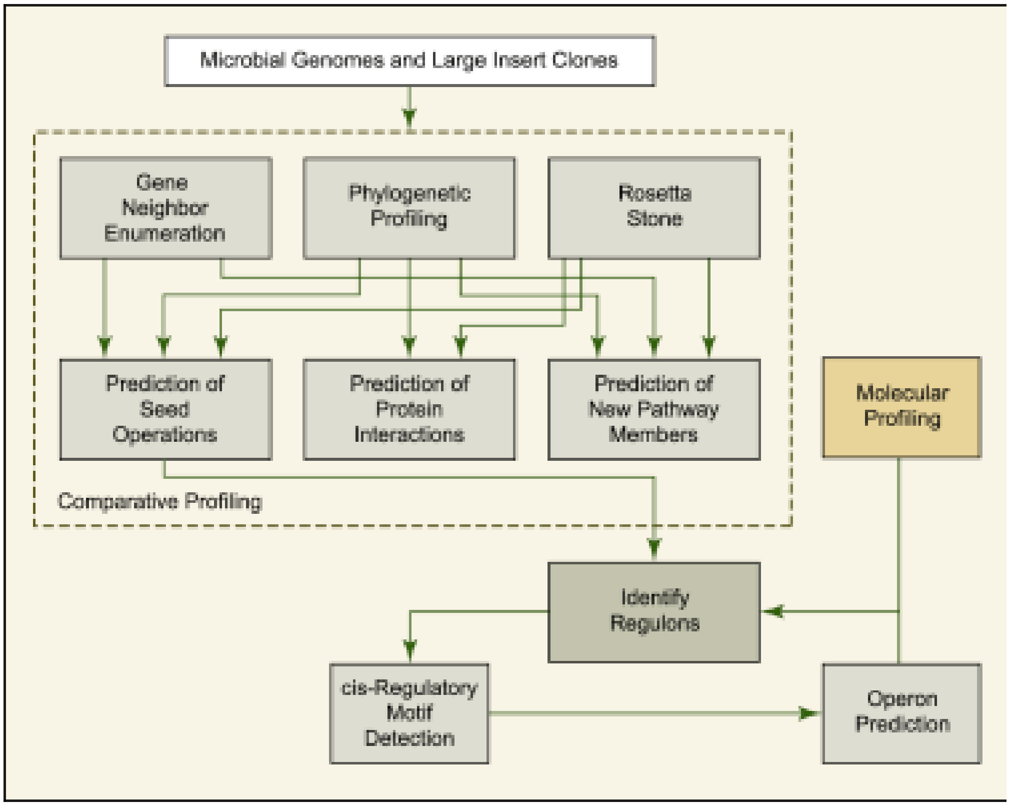

Microbial Genomes and Large Insert Clones
Gene Neighbor Enumeration
Phylogenetic Profiling
Rosetta Stone
Prediction of Seed Operations
Prediction of Protein Interactions
Prediction of New Pathway Members
Molecular Profiling
Comparative Profiling
Identify Regulons
cis-Regulatory Motif Detection
Operon Prediction

Figure 4.13. The regulon prediction pipeline.

hypothesis of which genes are co-expressed. Further, genes found in the same operon in one organism may not be found in that same operon in another. However, in this other organism, different proteins might be the target protein operons. It is likely that these other proteins might also be involved in the process. In this way, cross comparisons can lead to an expanded set of proteins that might be important to consider when creating a pathway model. Comparative analysis, as shown above, can also lead to predictions of regulatory sequence. Here we propose to set up a comparative genomics pipeline that will combine the data from the multiple microbial genome projects and the large insert clones obtained by the FGC, to predict the operon structure of the target organisms and some of their surrounding community and to yield hypotheses on the protein participants in stress response and their cis-regulatory features.

In this discussion we assume that the three target organisms will be annotated in the first year by the consortiums that sequenced them. In the mean time, we have begun to perform blast comparisons among all the microbial genomes with our targets (*G. metallireducens* is not available yet, however) to get an initial set of predicted genes. We will refine this list as new information becomes available.

Phylogenetic Profiles and Operon Prediction. As mentioned in the overview, one task will be to compile and link together the set of comparative computational tools that have recently become available for predicting cis-regulatory sequence and protein-protein interactions from cross-comparison of many microbial genomes. Figure 4.13 summarizes the pipeline to be developed and each box in that diagram is explained below. We have already installed all these tools on-site. The goal in this section is to link them together in an automated analysis pipeline such that as new genomes become available from external sources and as large-insert clones are produced by the AEMC and FGC, the predictions of this pipeline can be updated and served back to the investigators. The core technologies on which this pipeline is to be built are developed in other laboratories, but have proven themselves sufficiently useful to be incorporated into this effort. The core technologies are:

Gene Neighbor and Operon Estimation [27, 29, 34]. In these methods, the physical proximity of genes on the genome is used as a marker of functional relatedness and probability that they are co-regulated. If two proteins are found proximal to one another in a wide evolutionary diversity of microorganisms it seems to be a good bet that they are in the same operon and further that they are in the same pathway. Of course, both intracellular recombination processes and lateral transfers ensure that two proteins in an operon in one organism may be in different ones in a second. However, the new proteins nearby the member of this split-pair in this second organism may be new proteins in a regulon if they are also found in the original organism. Finally, all the noncoding regions surrounding the discovered protein sets is likely to contain the cis-regulatory regions that control their expression. Both probabilistic segmentation tools [30] and alignment tools like those discussed below and in reference [33] can aid in discovering these elements from this information.

Phylogenetic Profiling [37]. In this work, Pellegrini et al., develop a method based on assigning a vector to each protein in a starting each element of which indicates whether or not this protein is present or absent in a given organism. They show that proteins with similar vectors tend to be functionally linked; they show up in the same pathway or in the same structural complex. Proteins may be clustered using this profile and functional groups (and annotations) may be made based on this clustering. Thus, unannotated genes may be discovered to be part of a known pathway, and new interactions among proteins may be discovered. The latter, is aided by not just by profiling proteins but predicted domains as well (Rosetta Stone Method) [36]. Again proteins with similar profiles may be transcriptionally co-regulated so intergenic sequence drawn proximal to these proteins may be collected and aligned as above to discover cis-regulatory regions. It is in this task that Jonathan Eisen's expertise becomes essential. Determining the homologies among proteins in divergent phylogenetic classes is his specialty [38, 48]. Thus, he will aid in the robust annotation of the target genomes and the clones for use in this analysis.

Co-regulation Clustering to Propose Functional Interactions and Elements. This technology is based on combining molecular profiling technology and sequencing. mRNA arrays and proteomics technologies yield a measure of the changes in concentration of many species involved in functional response. These changes may be controlled at the transcriptional or translational level. In either case, the sequence determinants of these control elements are found in the open-reading frame and the surrounding noncoding regions for the genes of these species.

When genes and proteins exhibiting similar response profiles are clustered then the underlying sequences may be aligned and regulatory motifs determined [31, 35].

However, using these tools is not as simple as downloading them or recoding them and wrapping them in a CGI script. All these methods have parameters that score similarity, or cluster data in different ways. Each set of parameters will have different false-positive and false-negative rates that can be estimated from boot-strapping analyses using known regulatory sites and interactions. We will tune these parameters by compiling a list of the stress response genes from the annotated complete microbial genomes and the biochemically characterized cis-regulatory regions in *E. coli* and *B. subtilis* [lists of which can be found in such places as RegulonDB[49]or Subtilist[50]]. Once a tuned and (semi-) automated scoring of likely cis-regulatory regions, new pathway members, and putative protein-protein interactions can be made, then these will be applied to our target organisms and the large insert cloning set. Prediction of the cis-regulatory elements will be sent back to the FGC for validation by both knock-out and by attempting to affinity purify proteins attached to these regions [51]. New pathway member predictions will be sent to the knock-out core in the FGC for prioritization and interaction predictions will be sent to the FGC interaction cores for validation.

Regulatory Sequence Detection by Alignment. We will use modification of VISTA in order to find cis-regulatory regions that control expression of the stress response genes. In previous work, when two orthologous sequences are aligned, the analysis of the DNA sequence conservation could be performed and a position of every conserved region could be defined based on the annotation associated with a particular bacterial sequence. Our conservation analysis was aimed at identifying highly conserved regions with a cutoff of at least 70% identity (%ID) over 100 bp DNA region. This definition of cutoff is obviously an arbitrary one, but this simple approximation allowed retrieval of important regulatory sequences in mammals [52]. However we will also address the issue of much shorter and highly conserved DNA elements which are potentially biologically significant and can be currently missed while using the standard 70%/100bp conservation criteria.

To expand the definition for sequence conservation we will analyze regions as short as 20bp. The probability of finding an identical 20mer in two random sequences of 50 kb is less than 0.25%, while the probability of finding

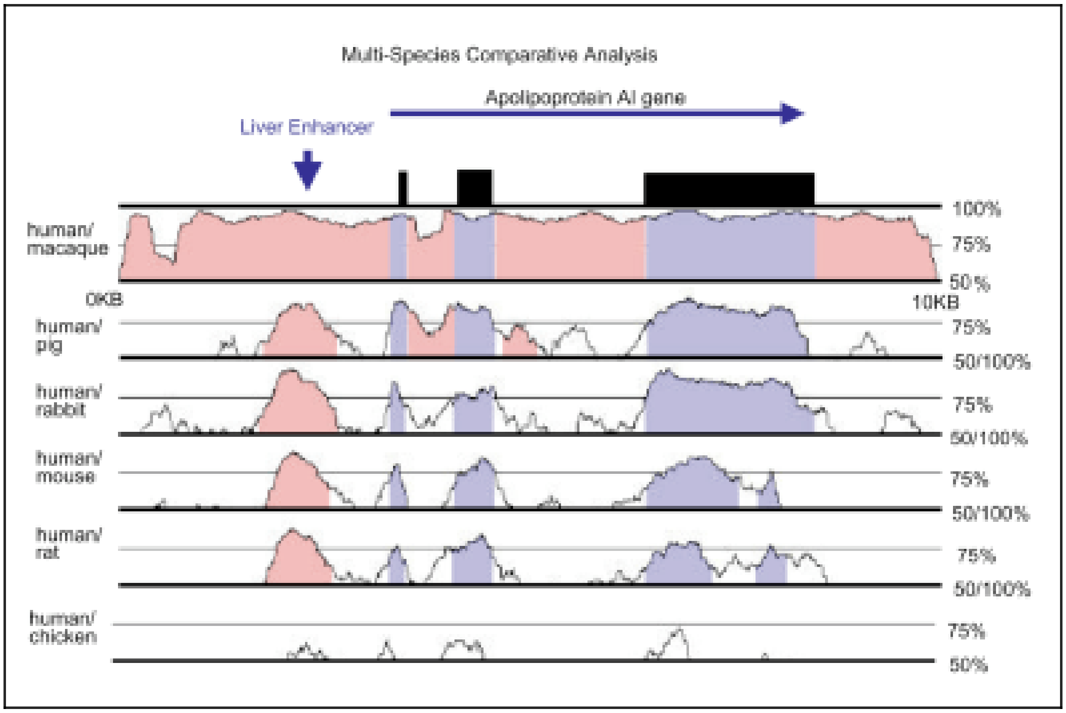


Figure 4.14. Multivista alignment of many species against human.

a 20 mer identical match in a random global alignment of the same size drops to less than $5*10^{-8}$. This simple calculation suggests that regions of 100%/20 bp have a small probability of being randomly conserved. We will also consider conserved regions shorter than 100% identity and shorter than 100bp in length using a sliding scale of sequence identity including: 100.0%/20 bp; 95.0%/30 bp; 90.0%/40 bp; 85.0%/50 bp; 80.0%/80 bp; 75.0%/90 bp; and 70.0%/100 bp.

After analyzing the alignment and retrieving conserved noncoding sequences that potentially can play regulatory roles, the question therefore remains as to whether most of these noncoding sequences are conserved due to functional constraints or are the result of a lack of divergence time? Based on the supposition that actively conserved noncoding sequences will be present in a third bacteria while noncoding regions that are similar because of an insufficient accumulation of random mutations will be absent, we have developed an algorithm to search for blocks of similarity in globally aligned sequences[32]. A problem in using sequence comparison algorithms designed to identify conserved coding sequences is that functionally conserved noncoding sequences frequently contain small insertions and deletions.

In previous work, to establish the cutoff criteria (X% identity over Y bp) for defining noncoding sequences as actively conserved we examined the three two-way sequence alignments [human/dog (H/D), human/mouse (H/M) and mouse/dog (M/D)] using intersection/union (I/U) analyses. The criteria for which these alignments had the largest number of overlapping and the least number of unique conserved noncoding elements were found to be different for three pairs of species: H/D (≥ 88% identity over ≥ 120 bp), H/M (≥ 80% identity over ≥ 120 bp), and D/M (≥ 75% identity over ≥ 120 bp).

The same approach will be used in the pairwise analysis of *D. vulgaris, S. oneidensis*, and *G. metallireducens*. We will evaluate three pairwise cutoffs of active conservation in order to further use them in analyzing conserved noncoding elements in search of their cis-regulatory function.

MultiVista allows for comparison of sequences of multiple organisms and visualizing them on the same scale (Figure 4.14). This example demonstrates how sequences can be sorted on their similarity to a base sequence (in this case a human genomic interval). Using each of the three sequences (*D. vulgaris, S. oneidensis*, and *G. metallireducens*) as a base we can sequentially align sequences of other organisms from the TIGR database and homologous regions from nearby bacteria obtained by large insert cloning and analyze levels of conservation in different regions of base genomes. Cutoffs for active sequence conservation will be established for all combinations of organisms in automatic manner.

Clustering by Regulatory Strategy. One of the central goals of this proposal is to discover the functional differences among the homologous stress response pathways both in the target organisms and in their immediate community (as sampled by the large insert cloning core). Further, it is a goal to determine if similar regulatory structures tend to aggregate within niches (e.g., anerobes vs. microaerophile) and if there is even a further stratification of strategies within a niche that indicated competition or cooperation among then local community. Another way to phrase this is we wish to use pathway elements (proteins and their domains, molecular interactions, cis-regulatory regions) as predictors of the functional class of the bacterium or where it lives. This is essentially and classification or discrimination task.

One approach to this task is to develop a variant of the phylogenetic profiling discussed above in which every molecule, regulatory element and interaction is scored as being present or absent (or unknown) in every sequenced microorganism (or large-insert clone). Of course, there are some subtleties to choosing what to include in the list of pathway elements. For example, an interaction cannot exist if one or both of the participating molecules do not exist in the organisms. There may be a caveat even to this observation if there has been domain shuffling, and the interaction domain is attached to another protein. This is one of the places where access to a graph database becomes very useful because to construct non-redundant elements subgraphs of molecules and their interactions might appear as single units. The ability to search the database with graph queries (based on one organism) to different levels of specification in another will be extremely useful. Similar clustering algorithms to those used in the standard phylogenetic clustering can be applied here to gain an estimate of the number of different strategies. (Clusters can be made both of regulatory elements by phylogenetic vector, or by organism by common elements.) By labeling the cells with other information such as their metabolic biases or niche specificities, or by taxonomic class, estimates can be made about which elements best predict those classifications (more less complicated classifiers can be applied to this task.

However, it is not yet clear which types of features to include in this analysis. We will have to try a number of different approaches before settling on a choice. We propose to employ a model for determining which pathway features most determine a bacterial function label (like aerobe) by adapting a variational inference model first developed for determining which symptoms most determine a disease. This model is called the Quick Medical Reference-Decision Theoretic (QMR-DT) method and it is far more general than this application [53]. Indeed, the basic model is similar the dynamic Bayes models discussed below. In our case, the "symptoms" are pathway features, the vector of this is denoted by ***f***, and the "diseases" are the classifications of the organism, denoted by ***l***. The elements of ***f*** and ***l*** are binary (either the feature exists or it doesn't, for example), and so, *a priori* ***f*** and ***l*** are binary random variables. The elements of these vectors form two layers of nodes in the graphical model, where the ***l*** nodes impinge on the ***f*** nodes. Thus, we try to estimate the probability of finding a particular pattern of features given a series of labels. Jordan et al. give a formal mathematical solution to estimation of this distribution given a set of data just like this one. Once the algorithm converges, one of the results is a set of weights that represent how much a particular label determines the presence of a feature. We propose to interpret these weights as indicators of which pathway features are most necessary for achieving the property indicated by the label—so for example, which feature most confers the ability to grow microaerophilically.

The results of this analysis will be used during the comparative modeling tasks of this proposal as described below. Further, the list of features most discriminatory for a particular type of bacterium is a prediction of elements that hypothetically should be found in all members of that type. These features (molecular and interactions) can be sent back to the FGC for validation in the target organisms. One interesting off-shoot of this analysis (with the modeling analysis below) may be the discovery of regulatory motifs (subnetworks) that are found multiple times across different, possibly evolutionarily distant, organisms. If there is any modularity to biological systems, these motifs may indicate where module boundaries exist.

4.4 Pathway Deduction

4.4.1 Overview

Once a set of reliable data has been collected and scored according the techniques outlined in Section 4.3, then pathway deduction can begin. Some of the resultant network hypothesis is derived immediately from the direct interaction data such as the phage display traps and the protein-crosslink mass spectrometry. The molecular complement that will be used to construct the full network hypothesis is defined as that set of molecules whose dynamics are changed under different stress conditions. These molecular elements include, of course, the proteins, mRNA, metabolites and chemical/physical stressors. However, the predicted regulatory elements are also members of this list. Starting with the list of putative interactions and molecular elements, the rest of the network must be deduced indirectly from the patterns of expression and activity of the measured species under different perturbations.

Conceptually, this task is the same as the standard engineering method of system identification. This involves a process whereby a physical system and a model of that system are both perturbed by the same signal. The physical system is perturbed by a physical signal such as a chemical or a temperature shift. The model is perturbed by introducing changes in a cognate variable. If measured output of the physical system and the calculated output of the model do not match over these perturbations then the physical model is tuned to better match it. Standard methods work relatively well for linear systems but for non-linear systems with memory, such a framework will, theoretically fail. It is, after all, a brand of the Halting Problem.

However, most of these identification techniques assume no physical knowledge of the unknown system. Here, we will have a good deal of prior knowledge of these systems: an estimate of the number of molecular players, and a list of many of their interactions (to some confidence). Most system identification methods also assume the system is linear and largely deterministic. However, in cellular systems both these precepts are, in general false.

Further, it is common that the system model is "unstructured"—meaning that it is composed simply of a generic regression model of specified order. Such models are not as much use as more structured causal model that can be validated by experiment directly (since the variables generally refer to physical entities that can be manipulated.) Nonetheless, we can borrow a great deal from this field and related work.

4.4.2 Background and Significance

The precursor to any model is data. The minimal basis of a cellular network model is a list of the molecular players and a list of the "influences" of one set of players on another or on a lumped cell behavior (such as growth). Molecular players and their interactions have traditionally been discovered through painstaking genetic and biochemical experiment. Experimental techniques, like those mentioned above for two hybrid interaction, or other techniques such as FRET, or protein cross-link mass-spectrometry can yield direct pathway information.

Newer experimental technologies have been developed for indirectly deducing interactions among molecules. These include: the multiple alignment of DNA regions upstream of genes clustered by their expression patterns [35], statistical analysis of concomitant variation and temporal sequencing [54], and perturbation/response modeling [55-60]. There is also a whole host of "inference" methods that try to deduce network or even model structure using directed knockout, or designed perturbation studies [56, 61-63]. Most of these assume a Boolean structure to the network that is only very rarely appropriate [64, 65]. We will use the least structured of the perturbation response methods in order to avoid unnecessary assumptions on the network physics and structure. We have experience in such methods as described below.

4.4.3 Preliminary Results

Arkin, et al. have worked on a number of theoretical and experimental approaches to reverse engineering pathways through perturbation response designs like that diagramed in Figure 4.1. The simplest of these methods is called Correlation Metric Construction (CMC) [58]. In this method, a prediction of the reaction pathway is deduced from time-lagged correlation functions of two chemical species at a time, obtained from time-resolved concentration measurements of the system under noisy perturbation of a number of the target metabolite concentrations. These functions are converted into interspecies distances, which are then used in the construction of a multidimensional object; a specified two-dimensional projection together with a clustering algorithm yields the reaction pathway of the reacting system. Arkin et al. [59] report an experimental test of the theory on a part of an in vitro glycolysis system containing 8 enzymes and 14 metabolites. This work involved developing a metabolic profiling technique using capillary electrophoresis and constructing a single continuously stirred tank reactor system for obtain reproducible data for dynamical concentration measurements of the enzymatic network. Figure 4.15 shows the time series obtained from this apparatus. Figure 4.16 is the aforementioned correlation function. This is the core of the method in which the degree to which two species are correlated is a measure of their coupling in the pathway, and the asymmetry of the correlation decay in time indicates which direction signals travel in the network. From this diagram a network hypothesis can be constructed using multidimensional scaling, a heuristic on a hierarchical clustering algorithm and a bit of expert chemical knowledge (to resolve impossible reactions by following conservation of composition). Figure 4.17 shows the results of this analysis from the glycolytic network experiments in which the network hypothesis and the actual known network are very similar.

Two other information theoretical techniques for measuring the relatedness among species is proposed in [66] and shown to be superior to CMC in that that can measure the non-linear and multivariate relationships among species with the concomitant increase in the necessary number of data points to form a good network hypothesis. Missing variable and lack of time resolution also present difficulties for these methods (as for all methods) but as long as the system is not allow to reach a steady-state between perturbation then the method will yield a network hypothesis. Recently, still another, more deterministic, impulse-response based network deduction method was

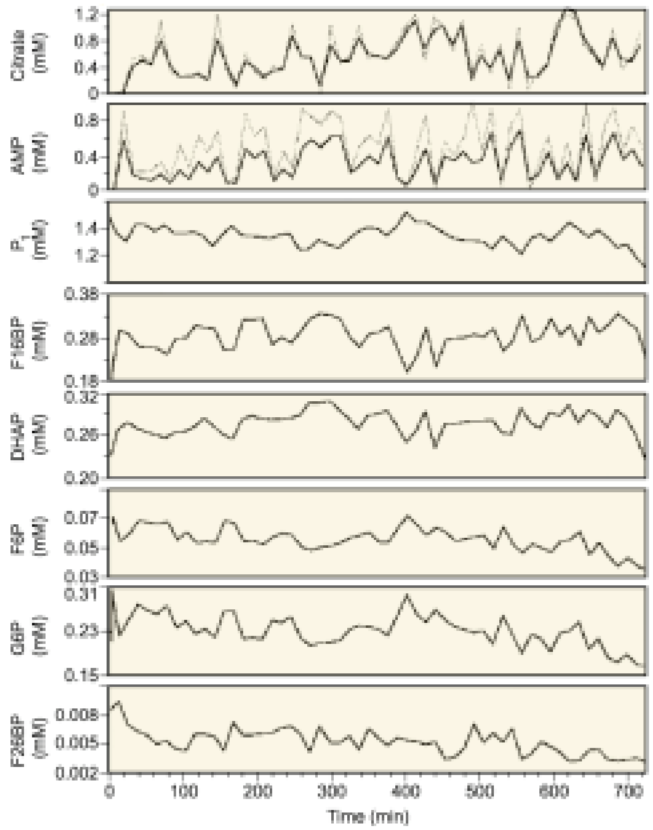


Figure 4.15. Time series of chemical concentration for the glycolytic system under perturbations in citrate and AMP concentration.

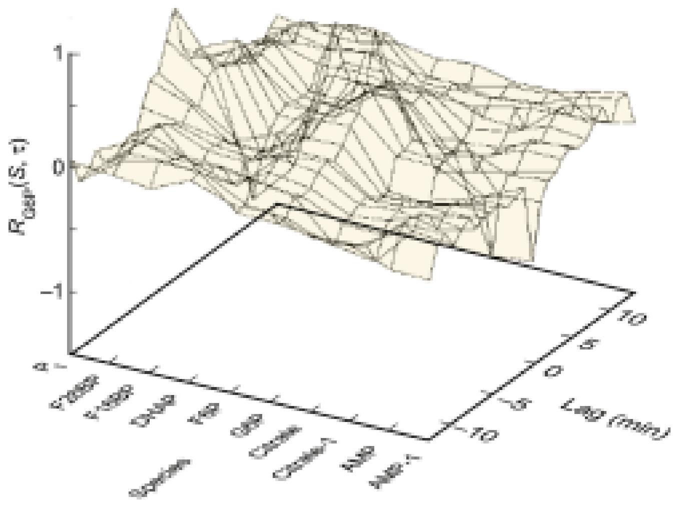


Figure 4.16. The correlation between Glucose-6-phosphate and all other species (in time).

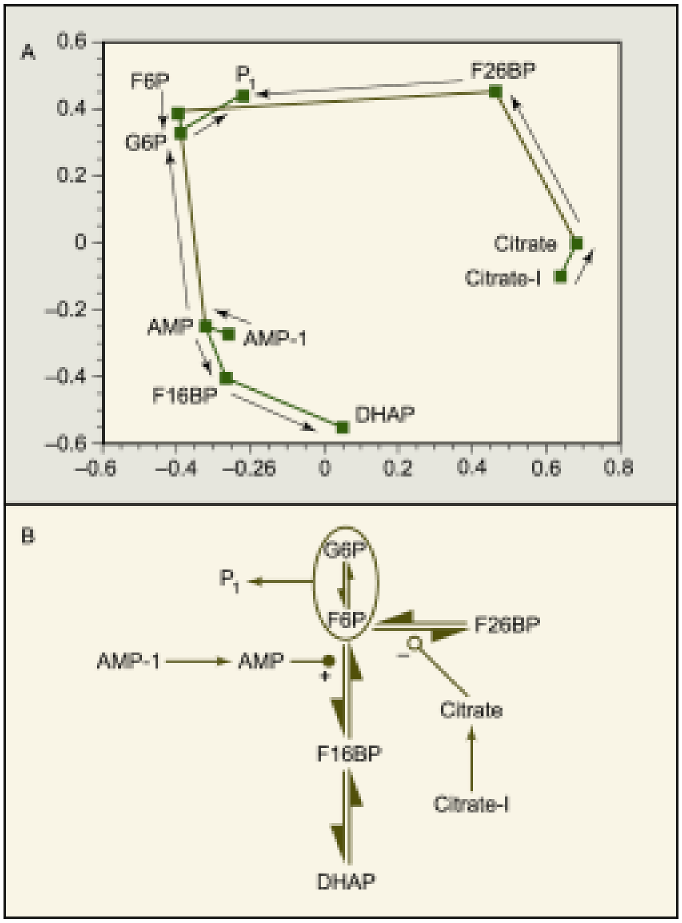


Figure 4.17. (a) The multidimensional scaling-based network hypothesis, and (b) the expert pathway diagram automatically deduced from (a).

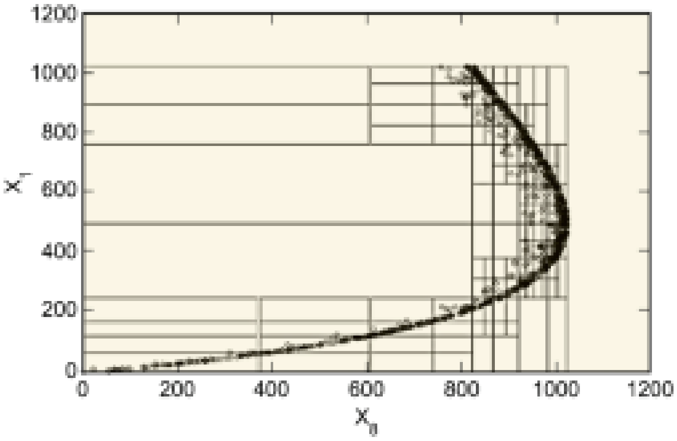


Figure 4.18. A measured relationship between two chemical species in an abstract reaction network driven by noise. The grid is made by a partitioning algorithm that calculates the mutual information [67].

proposed by Vance, Arkin, and Ross [68]. However, none of these approaches have yet been tested experimentally. Also, since many of these approaches require many "random" perturbations to be made to the system to fully probe its dynamics it would be useful to have criteria for exactly how to choose which perturbations would yield maximum information about the network.

4.4.4 Research Design and Methods

The methods described in preliminary results are especially useful when all you have is molecular profiling data and no external knowledge of the interactions among the species. Further, they are not easily scorable against the data: how much of the data are explained by the hypothetical network structure? How stable are these results to missing information or to noise on the measurements? The FGC, aside from the molecular profiling data, will also be giving us data on the direct interactions among some of the components thus it will be important to be able to incorporate this information into the pathway hypothesis data. Interaction hypotheses will also becoming from the comparative analyses above. Given the black-box reverse engineering (BBRE) methods above and these data, it will also be important for assigning a score for each hypothesis against the available molecular profiling information for each organisms (remember that each organism may have a different network structure).

The proposed research is based on applying the aforementioned BBRE methods to learn the structure of some networks. We then propose to combine these methods with Dynamic Bayesian Network approaches that have recently been applied in other laboratories in order to come up with principled rankings for how well different network hypotheses explain the observed molecular profiling data. However, we begin by a very short discussion of experimental design considerations

4.4.4.1 Experimental Design Considerations. One of the more important issues we will have to face in collecting molecular profiling information is exactly what perturbations to make? How many different pH values should we use? Can we gain more information by doing two perturbations at once? What is the functional difference for network deduction in perturbing an input to the system such a temperature versus changing the network structure by a knockout? What knockout would best help to distinguish between alternative network hypotheses? All of these fall in the realm of experimental designs and are, in general, difficult questions to answer. However, there are some very basic technologies that can aid in answering these questions. For example, standard factorial and Latin square designs are powerful methods (for certain types of systems) for reducing the total number of experiments necessary when testing the effects of multiple perturbations [69]. Thus, when the methods below call for "random" perturbations, these will be specially chosen random sequences that are permuted in such as way as to be consistent with such designs.

Newer techniques that mainly apply to choosing experiments the best reveal the structure of Boolean networks may also have application in these experiments [56]. Although these networks have decidedly non-Boolean character, the knock-out experiments are tantamount to the logical turn "on" and "off" of particular network capabilities. Genetic knockouts remove regions of DNA, transcripts and proteins from the network (affecting possibly many paths through the network at once) and chemical "knock-outs" precisely target particular interactions (or edges) in the network. Ideker et al. [56] developed a "predictor/chooser" method that, from a set of "knockout" perturbations (on a Boolean network), predicts a set of Boolean networks consistent with the existing data, and chooses the next knockout to be perform to maximally distinguish among network possibilities. In our case, although the knockouts are roughly Boolean in nature, the network response is not, nor are all perturbations knockouts. Thus, this method would have to be modified to be applicable. We will work on making such modifications.

4.4.4.1 Black-Box and Bayesian Methods. Of the three methods outlined in Samoilov and Ross [66], CMC, Entropy Metric Construction (EMC) and the Entropy Reduction Method (ERM), only CMC has been experimentally tested. The data requirements for ERM are beyond the abilities of the FGC as proposed. CMC is a useful procedure and will certainly be applied to the data as it comes off the pipeline. However, there are certain strong limitations to CMC that need to be addressed. For example, many interactions among pathway species are nonlinear. If there is a nonlinear but monotonic relationship between the value of one variable in the network and the value of another then CMC still works based on a non-parametric rank correlation. However, in cases where the state of one variable is a nonmontonic function of another then CMC will outright fail. Nonmonotonic functions are common in biology: for example, the activity of the □-phage P_{RM} promoter as a function of increasing repressor concentration first increases then decreases [70] similarly with the activity of the type-1 pili phase variation switch as a function of active leucine responsive protein [71], or the activity of MAP Kinase as a function of a scaffolding protein [72]. As an abstract example, consider the data shown in Figure 4.18. These data was derived from an abstract model of a chemical network, in which one of the species X_1 is an activator at low concentrations and inhibitor at high. A linear correlation would miss the relationship between X_1 and X_8. However, a measure based on the mutual information easily finds this relationship and is even extensible to more than one variable. We have tested the method on synthetic data such as this in which a hypothetical chemical reaction network is perturbed by external signals (chemical concentrations) as a function of time and had excellent results in reproducing the underlying network structure. However, it remains to be seen if it will work with real data. We will do that here. In any case, once a network hypothesis is forwarded, both by simply asserting a subnetwork using the experimental interaction data in union with the network predicting using EMC, we still need a method for asking how well this network explains the data. Hartemink et al. [73], propose a scoring scheme based on a Bayes Network comparison of a network hypothesis to data (this technique is very much related to the variational method of QMR-DT discussed above). In this technique, molecular profiling data are quantified (perhaps by simply scoring every measurement as significantly different from a control sample or not), and these data are used to parameterize an observed node in the underlying graphical network, protein concentrations are latent variables. The directionality and labels on the arrows among the observed and latent variables comprise the network hypothesis. Using this framework, the relative likelihood that observed data is generated by the proposed network may be calculated. Different network hypotheses can be ranked by this method. We will apply this approach to our data. However, in our case, we are not working only with microarray data so that the structure of the graphical model will be more complex. However, the basic theory of the method is not affected by inclusion of extra nodes for our multiple molecular data types.

Interactions proposed by high ranking network hypotheses can then be sent back to the FGC so that the molecular interaction cores for verification. One result of the Bayes analysis is that the relative contribution of each experiment to the explanatory power of the network hypotheses can be made. High ranked interactions, assuming they are proposed by the indirect methods (e.g., the BBRE methods), can then be sent back to the genetic and chemical knockout cores to be further probed.

4.5 Models of Pathway Function

4.5.1 Overview

Synthesis of data into formal models of cellular function is rapidly becoming a necessary industry. The complexity of the interactions among cellular constituents and the quantity of data about these interactions hinders the ability to predict how cells will respond to perturbation and how they can be engineered for industrial or medical purposes. Models provide a systematic framework to describe and analyze these complex systems. In the last few years, models have begun to have an impact on mainstream biology by creating deeper insight into the design rules of cellular signal processing, providing a basis for rational engineering of cells, and for resolving debates about the root causes of certain cellular behaviors. In fact, Arkin recently enumerated five reasons to construct a model of a system:

- **To demonstrate a design property of a network**. Quite often a network is so complex or so odd in structure that it is of interest to understand what properties of its design are necessary for cellular function.
- **To develop an understanding of endogenous control**. When more mechanistic models are available precise statements about control in the networks can be made.

- **To develop a strategy for control or design**. Models can be used to test design ideas for engineering networks and network behaviors in cells.
- **To prove necessity and/or sufficiency.** Given an observed cell behavior, models can be used to prove necessity of a given regulatory motif or sufficiency of known interactions to produce the phenomenon.
- **To explain the contradictory or exotic behavior of complex networks.** Contradictions in the literature and data about any particular pathway abound. Models can help test the different implications.

The second and third of these are the focus of this work as we try and construct and understanding of the stress response pathways in our key organisms. One of the more important reasons to make a model is simply to store, in the clearest possible format, all the hypotheses about who interacts with whom in a complex causal system. Let us take response to oxygen stress in *E. coli* as an example since a related stress response pathway will be examined in *S. oneidensis, G. metallireducens,* and *D. vulgaris*. Storz and Imlay [74] outline the basics of this pathway in which oxygen diffusing into the cell is converted into a free radical through the action of flavoproteins. The oxygen radicals can then damage the cell by interacting with sulfur-iron clusters such as that found in aconitase (one of the central biosynthetic enzymes). The released iron can aid in the formation of hydroxyl radicals than then attack DNA. As you might imagine these perturbations ultimately couple into nearly all the stress response pathways in *E. coli* including excision repair, osmotic stress, starvation, acid shock, stationary phase, and other pathways. Damaging DNA couples (randomly) to other gene expression systems, and the interaction with iron co-factors affects a predictable number of metabolic and biosynthetic pathways. Just the coupling to $\square_s$ expression has a vast effect of coupling these stress responses. Loewen et al. [75]enumerate 47 genes positively regulated by $\square_s$ and seven genes negatively regulated. The molecular players are metals, metabolites, small functional RNAs, mRNA, DNA regulatory regions and proteins including transcription factors, enzymes, kinases and transporters and transporters. There are a plethora of physical processes that must be considered as well, transport into and out of the cell (so coupling to other cells in the environment as well), gene expression, DNA damage, redox reactions, free radical reactions, small RNA control and more. To keep track of all this information about which molecules interact with which others, and the physical implications of the mechanisms of these interactions models are going to be necessary. The question is, even given a network hypothesis from the reverse engineering techniques in Subsection 4.6, what type of model should be built to achieve the goals stated above?

4.5.2 Background and Significance

4.5.2.1. Modeling Background. Because of the heterogeneity in data type, quality and availability, cellular network modelers have had to develop a number of different model classes that can operate at different levels of abstraction. The most common models are graphical models, i.e. cartoons, of the process. Cartoons graphically depict each biological component connected to others with arrows indicating their interaction. There is little standard nomenclature for cartoons although at least two formal graphical annotations have been suggested recently [76, 77]. However, even Kurt Kohn's wonderful summary cartoon of the mammalian cell cycle [76], wherein the conventions for representing interactions and species are outlined in some detail, contains abstractions and missing information. One example in that system is the representation of the important tumor suppressor (p. 53). The diagram shows this single protein having 27 sites to which phosphate can be added or removed by various specific enzymes. The implication is that this protein, theoretically, can be in any of 2^{27}=134,217,728 possible phosphorylation states and each of these state has a possibly different Gibbs free energy and different interaction kinetics with all other molecules in the system. This is a problem not only for this model type but spells trouble for any detailed mechanistic model of this system. Nevertheless, the graphical models summarizes a great deal of the current information about a pathway and facilitates the formation of hypotheses about network function as well as pointing out some of the difficulties involved in understanding the network.

Qualitative Models are the first automatically analyzable form of a model beyond a cartoon. They range from simple graphs to logical and statistical models. For example, Jeong et al. [78, 79] used only the yeast protein-protein interaction data cited above to conclude that the statistical properties of the graph implied a particular stability of network function to most “deletions” in the graph. This conclusion was strengthened by correlation of the number of interactions per protein with phenotypes of knockout mutants collected from the literature. For more dynamical and specific questions, logical models often are used when mechanistic data are lacking. Boolean, fuzzy logical or rule-based systems have been developed to approach simulation of complex networks. Thieffry and Thomas review their pioneering work in this field [80]. Many groups have used this paradigm for modeling genetic and developmental systems. Lee et al. used fuzzy logic (a generalization of Boolean logic) as a supplement to kinetic models to include uncertain information necessary for fitting the kinetics of metabolic enzymes [81]. Trelease et al. use a general qualitative simulation tool, QSIM, to simulate the effect of exogenous gene activations in the NF-□B network [82]. All these models require expert insight to codify the high-level rules in a consistent and accurate fashion. Because so much interpretation of the data is necessary before a model is made, there is an increased danger of building in a desired answer.

When perturbation response or time-series data are available, statistical influence models become feasible. Linear [83], neural network-like [84], and Bayesian models [54] all have been applied to deduce both the topology of gene expression networks and their dynamics. The amount of data necessary to fit these models often prohibits their use. Statistical influence models are not precisely causal models in that they are fits of model structure to indirect data on interactions. Interpretations of control in these models must be cautious.

With enough data, more mechanistic models can be developed. Cybernetic [85] and power law [86] formalisms assert a causal structure, but employ generic nonlinear functions formally parameterized by kinetic data and possibly constrained by optimality conditions. Such models form the basis of a large class of metabolic control analyses and dynamical simulations. More detailed models require that chemical or physical mechanisms be asserted for each interaction. For example, McAdams and Arkin, propose that, because of the small concentrations of the molecules involved, gene expression must be a stochastic process of a particular sort [87]. They follow the implications of the theory in an integrated model of the □-phage lysis/lysogeny decision and show that the decision is fundamentally nondeterministic [70]. Recently, there have even been stochastic models of stress response reported [88]. Physical models have the largest data requirement, are the most rigorously falsifiable, and, in principle, are the most predictive.

There is always a balance between top-down and bottom-up models. No model is fully bottom-up. Abstractions can both clarify the sources of control in a network and indicate where more data are necessary. There is always the problem of unknown players, and unknown and uncharacterized interactions in the network. A formal model that can be represented in mathematical form has the advantage of being a precise statement of the current understanding and can be formally proved or disproved and checked for consistency.

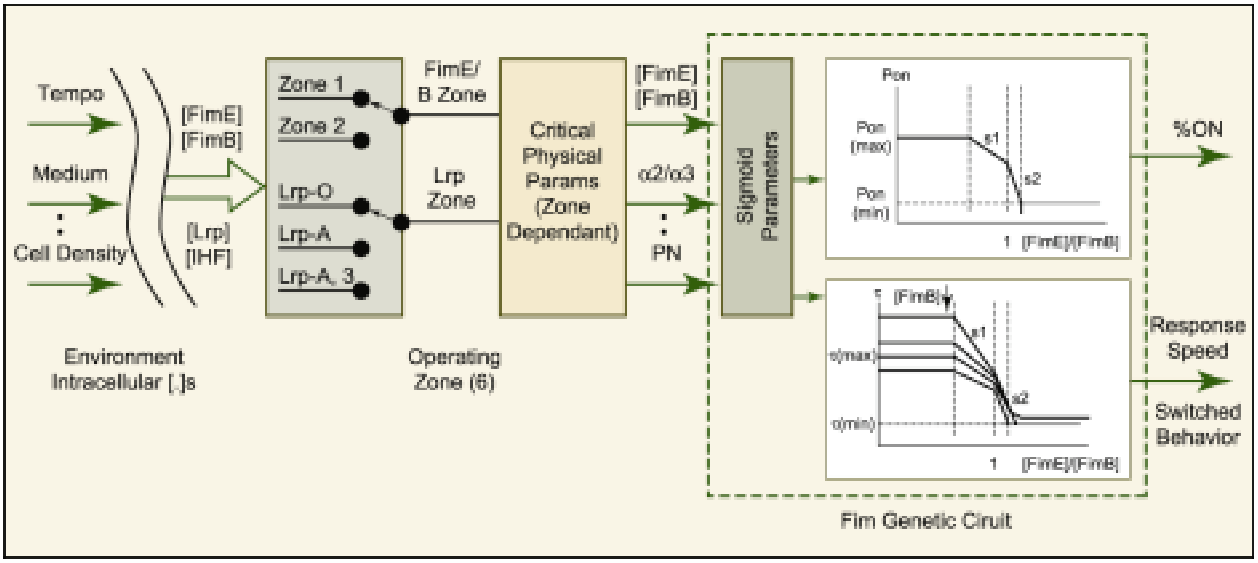


Figure 4.19. A summary of the control of the type-1 pili phase variation by external signals. This diagram is deduced from a physical model of all the underlying chemical and genetic network.

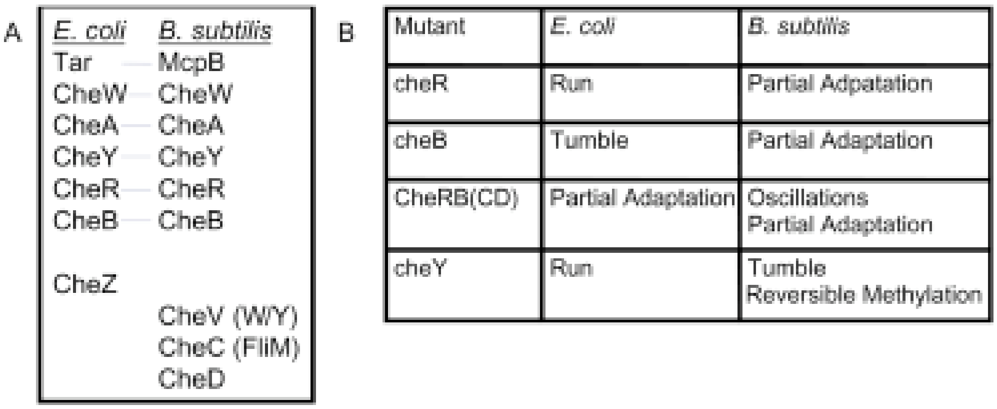
Almost all model types allow some sensitivity analysis to be performed. These vary from simple linear sensitivity or lyapunov analyses, through sophisticated bifurcation analyses, and response surface methodologies. Such analyses are important in determining how robust the predictions of the model are and identifying controllable elements.

4.5.3 Preliminary Results

The Arkin laboratory has long and varied experience in creating models of cellular processes. They include deterministic models of rat liver glycolysis and the TCA cycle that were analyzed for flux control and mode switching [89], models to discriminate different hypotheses for calcium control of the eukaryotic cell cycle [90], stochastic models of gene expression [87] that were used to explore the effects of noise on the □-phage lysis/lysogeny decision [70], and a model of control of the type-1 pili phase variation system (a target, by the way, of $\square_s$ control and downstream targets of $\square_s$ control) in uropathic *E. coli*. This last system is a model of a stress response pathway in which the population stochastic inverts a region of DNA in order to turn on and off pili it uses to adhere to its host. Although the pili allow adherence to the host, and confer the ability to invade, the pili are also targets of the immune system. Thus, there are competing requirements for having the pili and the population stochastically turns on and off the pili such that there is usually only about 1% of the population that is piliated at any moment. When a fairly large hybrid stochastic/thermodynamic model of this system was made, the model could be analyzed to uncover the modes of control that lead to two important population observables, the percent of the population piliated at any moment, and the rate at which this percentage could change given a change in conditions. Bacteria with different operating curves for these phenomena have different virulence indices such as persistence, recurrence, and transmissibility. A summary diagram is shown in Figure 4.19. Similar types of model will be developed here for stress response pathways based on the network hypotheses derived above.

Once models of one key organism has been made in deep detail and a control analysis has been accomplished so that it is clear what features of the network architecture are necessary for the observed behaviors, construction of models for homologous pathways both becomes easier and begins to address new questions. We have recently completed such a comparative analysis of the chemotactic pathways (another stress response system) in *E. coli* and *B. subtilis* [91]. There are hundreds of papers and many models of *E. coli* chemotaxis. There are less than twenty papers on the homologous pathway in *B. subtilis*. A simple phylogenetic profile for the two organisms yields Figure 4.20a. Nearly all the proteins in one system have homologs in the other. However, there are different numbers of receptor types in the two systems, and in the *che* operons there are a few different proteins. Moreover, up mutation the two systems behave differently as summarized in Figure 4.20b. Ultimately, by combining comparative analysis with control theoretical analyses we were able to predict the likely regulatory differences between the two organisms. Bioinformatic analyses yielded information on receptor methylation differences between the two organisms (that were also implied in methyl-group profiling experiments), the cheRB mutant data implied (after a control analysis) that *B. subtilis* had a cheY feedback directly to the receptor that *E. coli* does not possess (a conclusion that was backed up after the fact by a targeted comparative analysis that located a cheY binding domain homologous to the missing cheZ in the McpB receptor in *B. subtilis* and supported by point mutation data in that predicted region that show abolished feedback response) and these together lead to a formal hypothesis for the control of receptor activity by a novel methylation mechanism (completely different from *E. coli*) and a cheY feedback (Figure 1.9). The power of this analysis gives us confidence that similar findings will be made when comparing the three target organisms in this proposal. It also underscores the need for an integrated framework for bioinformatics and modeling.

To achieve this integration we have been developing a tool called BioSPICE that performs just this integration and allows models and data of the types discussed in this section to be developed in a “network bioinformatics” environment. This tool is far from a finished product and will have to be expanded and hardened to serve the needs of this proposal. The tools include a database (BioDB, see above), a set of bioinformatics and simulation tools and model development tools (One of the useful interfaces to the tool is shown in Figure 4.21. This is the pathway cartooning and display tool that can display molecular networks in various forms. Pathways may be brought up



A

E. coli	*B. subtilis*
Tar	McpB
CheW	CheW
CheA	CheA
CheY	CheY
CheR	CheR
CheB	CheB
CheZ	
	CheV (W/Y)
	CheC (FliM)
	CheD

B

Mutant	*E. coli*	*B. subtilis*
cheR	Run	Partial Adpatation
cheB	Tumble	Partial Adaptation
CheRB(CD)	Partial Adaptation	Oscillations Partial Adaptation
cheY	Run	Tumble Reversible Methylation

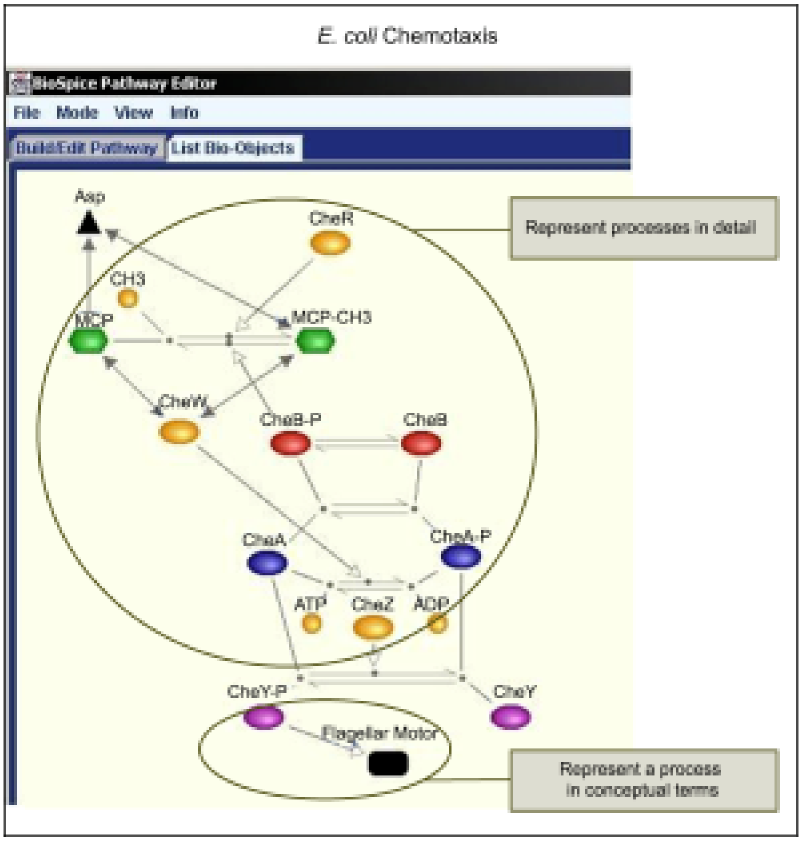
Figure 4.20. (a) Homologous proteins in the two bacteria and other protein in the *che* operons (*E. coli* homologies in parentheses. **(b)** Differences in chemotactic behaviors under different mutations.

from the database, decorating with physical information such as expression data, text annotated, and attached to models. There is also a suite of model analytic tools hooked in spanning third party bifurcation analysis tools [e.g., AUTO, [92]] and tools for response surface analysis [93]. This interface will be the primary entry point to allow the AEMC to interact with our simulators to develop their conceptual models.

4.5.4 Research Design and Methods

The main part of this research will be to actually develop the stress response models based on literature reading and the network hypotheses developed above. Inputs to the models will be time-dependent soil conditions, nutrient sources, metal concentrations and other environmental variables. The output from the models will be predicted molecular profiles, estimates of metal/radionuclide reduction rates and estimates of growth rates of the population. These will be compared to the data from the molecular profiling core, and the 16S RNA array data as well as the all the stress recovery data developed by the AEMC. The creation of this amount and diversity of data against which to validate models is a important aspect of this proposal. The data types and physical processes encountered in this project are different from those dealt with before in the BioSPICE modeling context. We will have to significantly expand the abilities of the tool to deal with phenomena like redox chemistry, DNA damage statistics, compartment models in which to implement cellular transport models, and mixed population models where many instantiations of the different models for our organisms are run together.

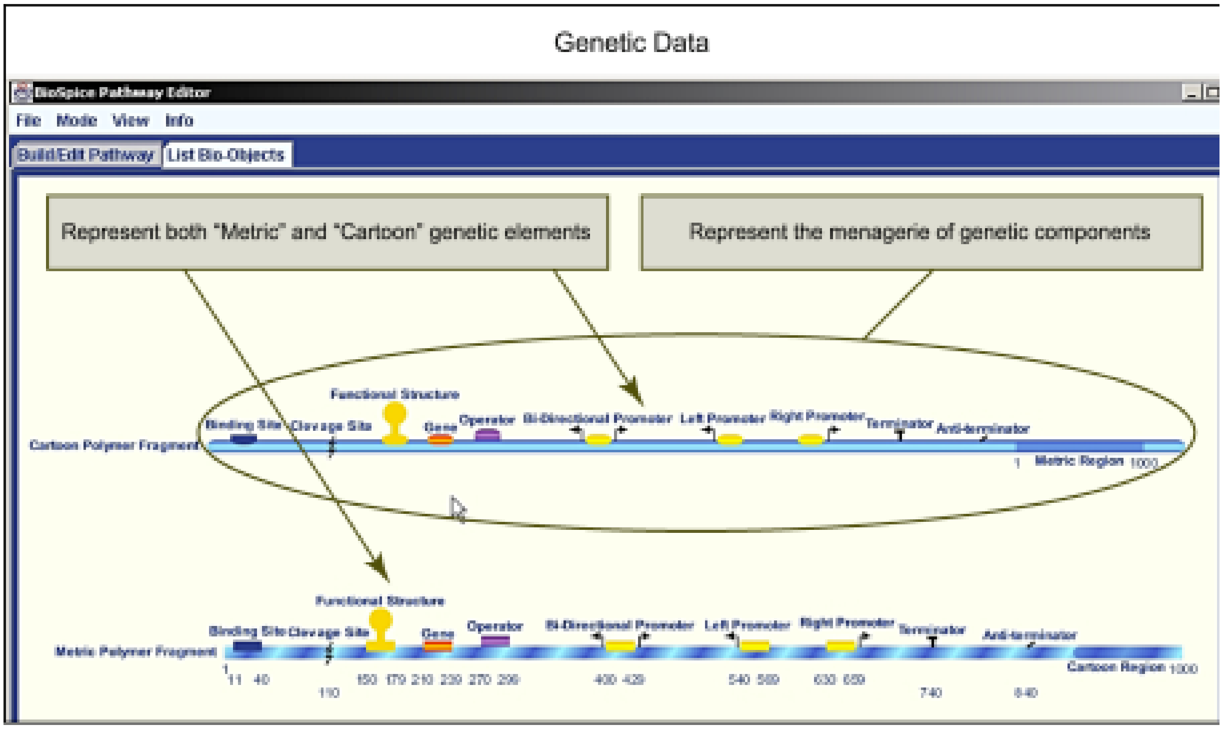
4.5.4.1. Modeling Stress Response. Currently we are modeling stress response in another soil microorganisms, *B. subtilis*. Here we are attacking the speculation pathways, competence, chemotaxis, various metabolisms, and a few other aspects. We expect to transfer many of the model aspects to our target organisms. However, we will begin by modeling the oxygen stress response pathways since these seem to be a critical route to metal/radionuclide reduction and an experimental condition that is easy to vary. We will begin by using the *E. coli* and *B. subtilis* oxygen stress pathways as a seed that we will run through the comparative genomics pipeline described above. From this we hope to create target set of molecules and regulatory regions that we will attempt to find in the genomes of our target organisms. A literature review will then be initiated to gather all the known information on interactions and mechanisms among these elements. The molecule and interaction list will be passed to the AEMC and FGC so they can prioritize their targets and measurements. The initial models will certainly be somewhat abstract ordinary differential equations expressing rough kinetics for each reaction set. Only if the flow cytometry data or FTIR imaging data require it will we move to the stochastic models that describe the endogenous noise in the system and predicts population heterogeneity.



Represent processes in detail

Represent a process in conceptual terms

Figure 4.21. (a) (left) A signal transduction network in the pathway builder showing different levels of modeling. The flagellar motor is a single functional unit for example. **(b) (below)** The interface allows sophisticated genetic information to be entered and worked with.



Represent both "Metric" and "Cartoon" genetic elements

Represent the menagerie of genetic components

As data and network hypotheses come out of the FGC and the network deduction tasks these models will be refined to incorporate those hypotheses. Models will be to one set of molecular profiling data and asked to then predict the outcome of other perturbation data and knockout experiments. Models can be scored by standard goodness of fit measures as well as the Bayesian measures discussed in Subsection 4.4.4.1. Comparative analyses like that described very briefly in the preliminary results will be brought to bear to construct and harden the models of the three target organisms. Once these models are validated against a fairly large number of data sets and knockout experiments, they may be used as a basis for “hypothetical” (meaning unvalidated) models of the response of the surrounding microbial community based on the modifying the regulatory network model using predictions from the regulatory feature clustering results from Subsection 4.3.2.4.3. The different regulatory feature sets essential reduce to changes in the model structure, each of the changes may lead to different utilizations of the external resources, activation of different combinations of stress response pathways, and different growth and reduction rates.

Once built these models will have to be combined together in one large model to ask how populations of these cells might compete for the external resources and how this would affect, finally, the total metal/radionuclide reducing power of the bacterial population. It is at this point that the AEMC scientists can truly engage in the conceptual model development to explore natural and biostimulatory attenuation of soil contaminants.

This is program of modeling is very ambitious and has not, to our knowledge, been attempted at this scale. Currently, to model just one pathway in one organism, typical hundreds of models at different levels of abstraction are built to test various ideas and hypotheses before the final model is valid enough to make predictions with. However, we have never had the diversity and amount of information that are planned for this project. It will be a central challenge to speed the model development cycle in the face of this type of data. Thus, we will rely heavily on the development of excellent data and model management tools and a streamlined data analytical framework to support model/data comparison and storage of results and hypotheses.

4.6 Experimental Core Facility: Computational Biology Facility (LBNL)

VIMSS will develop and support a unique computational facility for the systems biology of stress and survival pathways in bacteria. Three areas of service will be maintained: 1) the tools and data developed by the institute will be easily accessible to the project researchers and the extended scientific community for uploading, querying and analyzing the developed data; 2) the facility will be planned so that all software and data management tools may be applied to the expanding focus of VIMSS. For example, the institute team is already planning to scale-up to include stress response pathways in organisms considered to be threat agents. Stress pathways specific for interaction with animal hosts and the host environment must be expressible within the computational framework; 3) the data standards, exchange protocols, database schema, bioinformatics and model tools developed by the project will have application beyond the scope of the VIMSS. These should be made available to the scientific public in easy to install, well-documented form. Resources should be available to aid external researchers in implementing and using VIMSS software in their own projects. Similarly, VIMSS computational personnel will endeavor to adhere to external standards as closely as possible and still serve the needs of the project. Finally, the data in the VIMSS core must be well quality controlled, maintained and updated. A professional staff will manage the large set of tools to be developed, maintain the web-site, develop documentation, coordinate releases of software and data, and most importantly to curate and quality control the data and turn research code into robust software. The explicit definition of curation here includes the following responsibilities: 1) Checking data formats and compliance with data standards (both of experimental quality control and adherence to data exchange formats and included information, 2) Ensuring, 3) Making links of upload data to other data and information in the database and to external databases, 4) Revising and deprecating out-of-date information in the database, 5) Creating a quality rating metric for all information in the database and adding expert annotation when necessary, 6) developing and executing data release policies and formats. These needs and the goals of the VIMSS make the computational infrastructure no less of a user facility than any physical measurement facility such as a beamline.

The four named services of the facility are: 1) The Bacterial Stress and Survival Systems (B-SaSS) database (Section 4.2). 2) The Comparative Genomic Analysis of Stress and Survival Pathways (C-GASP) pipeline (Section 4.3.4.2). 3) The Data Analysis and Reverse Engineering for Stress and Survival Pathways (DARE-SSP) pipeline (Sections 4.2.4.1 and 4.4). 3) The Modeling of Stress and Survival Pathways (MoSSPath) environment (Section 4.5). Two other large efforts with sections complementary to the VIMSS Computational Core are currently funded: 1) the DARPA::BioSPICE project, developers of BioDB and tools applicable to *B. subtilis* sporulation and other stress response pathways. These are directly applicable to this project; 2) the Alliance for Cellular Signaling NIH GLUE grant, developers of data analysis and simulation of chemotaxis in B-cells. Some of these tools and data bases are similar in spirit to the ones described here. LBNL and UCB support joint programs in genomics and computational biology with over 20 principle investigators in this general area. LBNL run the National Energy Supercomputing Center with wide expertise in advanced computing, scientific data analysis, simulation and information management. The VIMSS staff will be part of this extended community of researchers and staff and be able to take advantage of their expertise and developments.

4.7 References

1. Karp, P.D., et al., *The EcoCyc and MetaCyc databases.* Nucleic Acids Res, 2000. **28**(1): p. 56-9.
2. Hull, R.B. and R. King, *Semantic database modelling: Survey, applications and research issues.* ACM Computing Surveys, 1987. **19**(3): p. 201-260.
3. Consens, M.P. and A.O. Mendelzon. *Graphlog: A Visual Formalism for Real Life Recursion*. in *Proceedings of ACM Principles of Database Systems Conference*. 1990.
4. Gyssens, M., et al., *A Graph-Oriented Object Database Model.* IEEE Transactions on Knowledge and Data Engineering, 1994. **6**(4): p. 572-586.
5. Ellis, G., *Compiled Hierarchical Retrieval,*, in *Conceptual Structures: Current Research and Practice*, P. Eklund, Editor. 1992, Ellis Horwood: New York.
6. Levinson, R.A., *Pattern associativity and the retrieval of semantic networks.* Computers & Mathematics with Applications Journal, 1992. **23**(2): p. 573-600.
7. Ehrig, H., Engels, G., Kreowski, H.J., and Rozenberg, G., *Applications, Languages, and Tools.* Handbook of Graph Grammars and Computing by Graph Transformation, ed. G. Engels, Kreowski, H.J., and Rozenberg, G. Vol. 2. 1998: World Scientific.
8. Rozenberg, G., *Foundations*. Handbook of Graph Grammars and Computing by Graph Transformation. Vol. 1. 1997, Singapore: World Scientific.
9. Dralyuk, I., et al., *ASDB: database of alternatively spliced genes.* Nucleic Acids Res, 2000. **28**(1): p. 296-7.
10. Markowitz, V.M. and F. Olken. *Schema design for a molecular biology laboratory information management system*. 1989.
11. Olken, F., et al. *Object lessons learned from a distributed system for remote building monitoring and operation*. in *1998 ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages and Applications (OOPSLA'98)*. 1998.
12. McEntire, R., et al., *An evaluation of ontology exchange languages for bioinformatics.* Proc Int Conf Intell Syst Mol Biol, 2000. **8**: p. 239-50.
13. Giaever, G., et al., *Functional profiling of the S. cerevisiae genome.* Nature, 2002: p. Submitted.
14. Ohlrich, M., et al. *SubGemini: Identifying Subcircuits Using a Fast Subgraph Isomorphism Algorithm*. in *Proceedings of the 30th IEEE/ACM Design Automation Conference*. 1993.
15. Rudolf, M. *Utilizing constraint satisfaction techniques for efficient graph pattern matching.* in *6th International Workshop on Theory an Application of Graph Transformations*. 1998.
16. van Helden, J., et al., *Representing and analysing molecular and cellular function using the computer.* Biol Chem, 2000. **381**(9-10): p. 921-35.
17. van Helden, J., et al., *From molecular activities and processes to biological function.* Brief Bioinform, 2001. **2**(1): p. 81-93.

18. Wang, X., et al., *Finding Patterns in Three Dimensional Graphs: Algorithms and Applications to Scientific Data Mining.* IEEE Transactions on Knowledge and Data Engineering, 2002: p. accepted.
19. Finkelstei, D., et al., *Microarray data quality analysis: lessons from the AFGC project. Arabidopsis Functional Genomics Consortium.* Plant Mol Biol, 2002. **48**(1-2): p. 119-31.
20. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.
21. Yang, Y.H., M.J. Buckley, and T.P. Speed, *Analysis of cDNA microarray images.* Brief Bioinform, 2001. **2**(4): p. 341-9.
22. Yang, Y.H., et al., *Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.* Nucleic Acids Res, 2002. **30**(4): p. e15.
23. Sherlock, G., *Analysis of large-scale gene expression data.* Brief Bioinform, 2001. **2**(4): p. 350-62.
24. Schwikowski, B., P. Uetz, and S. Fields, *A network of protein-protein interactions in yeast.* Nat Biotechnol, 2000. **18**(12): p. 1257-61.
25. Ito, T., et al., *Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins.* Proc Natl Acad Sci U S A, 2000. **97**(3): p. 1143-7.
26. Mrowka, R., A. Patzak, and H. Herzel, *Is there a bias in proteome research?* Genome Res, 2001. **11**(12): p. 1971-3.
27. Ermolaeva, M.D., O. White, and S.L. Salzberg, *Prediction of operons in microbial genomes.* Nucleic Acids Res, 2001. **29**(5): p. 1216-21.
28. Wolf, Y.I., et al., *Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context.* Genome Res, 2001. **11**(3): p. 356-72.
29. Overbeek, R., et al., *Use of contiguity on the chromosome to predict functional coupling.* In Silico Biol, 1999. **1**(2): p. 93-108.
30. Bussemaker, H.J., H. Li, and E.D. Siggia, *Regulatory element detection using a probabilistic segmentation model.* Proc Int Conf Intell Syst Mol Biol, 2000. **8**: p. 67-74.
31. Bussemaker, H.J., H. Li, and E.D. Siggia, *Regulatory element detection using correlation with expression.* Nat Genet, 2001. **27**(2): p. 167-71.
32. Dubchak, I., et al., *Active conservation of noncoding sequences revealed by three-way species comparisons.* Genome Res, 2000. **10**(9): p. 1304-6.
33. Hughes, J.D., et al., *Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae.* J Mol Biol, 2000. **296**(5): p. 1205-14.
34. Manson McGuire, A. and G.M. Church, *Predicting regulons and their cis-regulatory motifs by comparative genomics.* Nucleic Acids Res, 2000. **28**(22): p. 4523-30.
35. Tavazoie, S., et al., *Systematic determination of genetic network architecture.* Nat Genet, 1999. **22**(3): p. 281-5.
36. Marcotte, E.M., et al., *Detecting protein function and protein-protein interactions from genome sequences.* Science, 1999. **285**(5428): p. 751-3.
37. Pellegrini, M., et al., *Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.* Proc Natl Acad Sci U S A, 1999. **96**(8): p. 4285-8.
38. Pollock, D.D., et al., *A case for evolutionary genomics and the comprehensive examination of sequence biodiversity.* Mol Biol Evol, 2000. **17**(12): p. 1776-88.
39. Tatusov, R.L., E.V. Koonin, and D.J. Lipman, *A genomic perspective on protein families.* Science, 1997. **278**(5338): p. 631-7.
40. Florea, L., et al., *Web-based visualization tools for bacterial genome alignments.* Nucleic Acids Res, 2000. **28**(18): p. 3486-96.
41. Delcher, A.L., et al., *Alignment of whole genomes.* Nucleic Acids Res, 1999. **27**(11): p. 2369-76.
42. Schwartz, S., et al., *PipMaker--a web server for aligning two genomic DNA sequences.* Genome Res, 2000. **10**(4): p. 577-86.

43. Batzoglou, S., et al., *Human and mouse gene structure: comparative analysis and application to exon prediction.* Genome Res, 2000. **10**(7): p. 950-8.

44. Bray, N., et al., *AVID: A global alignment program.* 2002: p. in preparation.

45. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.

46. Kent, W.J., *BLAT---The BLAST-Like Alignment Tool*. 2002, Genome Research.

47. Pachter, L., *Domino Tiling, Gene Recognition and Mice*, in *Mathematics*. 1999, Massachusetts Institute of Technology: Cambridge.

48. Nierman, W.C., et al., *Genome data: what do we learn?* Curr Opin Struct Biol, 2000. **10**(3): p. 343-8.

49. Salgado, H., et al., *RegulonDB (version 3.2): transcriptional regulation and operon organization in Escherichia coli K-12.* Nucleic Acids Res, 2001. **29**(1): p. 72-4.

50. Moszer, I., et al., *SubtiList: the reference database for the Bacillus subtilis genome.* Nucleic Acids Res, 2002. **30**(1): p. 62-5.

51. McCue, L., et al., *Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes.* Nucleic Acids Res, 2001. **29**(3): p. 774-82.

52. Loots, G.G., et al., *Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons.* Science, 2000. **288**(5463): p. 136-40.

53. Jordan, M.I., et al., *An Introduction to variational methods for graphical models*, in *Learning in Graph Models*, M.I. Jordan, Editor. 1998, Kluwer Academic Publishers: Boston. p. 105-161.

54. Ben-Dor, A., et al., *Tissue classification with gene expression profiles.* Journal of Computational Biology, 2000. **7**(3-4): p. 559-83.

55. D'Haeseleer, P., S. Liang, and R. Somogyi, *Genetic network inference: from co-expression clustering to reverse engineering.* Bioinformatics, 2000. **16**(8): p. 707-26.

56. Ideker, T.E., V. Thorsson, and R.M. Karp, *Discovery of regulatory interactions through perturbation: inference and experimental design.* Pac Symp Biocomput, 2000: p. 305-16.

57. Ideker, T., et al., *Integrated genomic and proteomic analyses of a systematically perturbed metabolic network.* Science, 2001. **292**(5518): p. 929-34.

58. Arkin, A.P. and J. Ross, *Statistical Construction of Chemical Mechanisms from Measured Time-Series.* Journal of Physical Chemistry, 1995. **99**(3): p. 970-979.

59. Arkin, A.P., .Shen, P.-D., Ross, J., *A Test Case of Correlation Metric Construction of a Reaction Pathways from Measurements. Science*, 1997. **277**(5330): p. 1275.

60. Liang, S., S. Fuhrman, and R. Somogyi, *Reveal, a general reverse engineering algorithm for inference of genetic network architectures.* Pacific Symposium on Biocomputing, 1998. **95**(1): p. 18-29.

61. Maki, Y., et al., *Development of a system for the inference of large scale genetic networks.* Pac Symp Biocomput, 2001: p. 446-58.

62. Akutsu, T., S. Miyano, and S. Kuhara, *Algorithms for inferring qualitative models of biological networks.* Pacific Symposium on Biocomputing, 2000. **300**(5): p. 293-304.

63. Akutsu, T., S. Miyano, and S. Kuhara, *Inferring qualitative relations in genetic networks and metabolic pathways.* Bioinformatics, 2000. **16**(8): p. 727-34.

64. McAdams, H.H. and A. Arkin, *Simulation of prokaryotic genetic circuits.* Annu Rev Biophys Biomol Struct, 1998. **27**: p. 199-224.

65. Arkin, A., *Signal Processing by Biochemical Reaction Networks*, in *Self-Organized Biodynamics and Nonlinear Control*, J. Walleczek, Editor. 2000, Cambridge University Press: Cambridge. p. accepted.

66. Samoilov, M., A. Arkin, and J. Ross, *On the deduction of chemical reaction pathways from measurements of time series of concentrations.* Chaos, 2001. **11**(1): p. 108-14.

67. Fraser, A.M. and H.L. Swinney, *Independent coordinates for strange attractors from mutual information.* Phys. Rev. A, 1986. **33**: p. 1134-1140.

68. Vance, W., A.P. Arkin, and J. Ross, *Determination of causal connectivities of species in reaction networks.* Proceedings of the National Academy of Sciences, USA, 2002: p. accepted.

69. Hicks, C.R. and K.V. Turner, *Fundamental concepts in the design of experiments*. 5th ed. 1999, New York: Oxford University Press. x, 565.

70. Arkin, A., J. Ross, and H.H. McAdams, *Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells.* Genetics, 1998. **149**(4): p. 1633-48.

71. Wolf, D.M. and A.P. Arkin, *Fifteen Minutes of fim: Control of Type 1 Pili Expression in E. coli.* Omics, 2002. **6**(1): p. 91-115.

72. Levchenko, A., J. Bruck, and P.W. Sternberg, *Scaffold proteins may biphasically affect the levels of mitogen-activated protein kinase signaling and reduce its threshold properties.* Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(11): p. 5818-23.

73. Hartemink, A.J., et al., *Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks.* Pac Symp Biocomput, 2001: p. 422-33.

74. Storz, G. and J.A. Imlay, *Oxidative stress.* Curr Opin Microbiol, 1999. **2**(2): p. 188-94.

75. Loewen, P.C., et al., *Regulation in the rpoS regulon of Escherichia coli.* Can J Microbiol, 1998. **44**(8): p. 707-17.

76. Kohn, K.W., *Molecular interaction map of the mammalian cell cycle control and DNA repair systems.* Mol Biol Cell, 1999. **10**(8): p. 2703-34.

77. Pirson, I., et al., *The visual display of regulatory information and networks.* Trends Cell Biol, 2000. **10**(10): p. 404-8.

78. Jeong, H., et al., *Lethality and centrality in protein networks.* Nature, 2001. **411**(6833): p. 41-2.

79. Jeong, H., et al., *The large-scale organization of metabolic networks.* Nature, 2000. **407**(6804): p. 651-4.

80. Thieffry, D. and R. Thomas, *Qualitative analysis of gene networks.* Pac Symp Biocomput, 1998: p. 77-88.

81. Lee, B., et al., *Incorporating qualitative knowledge in enzyme kinetic models using fuzzy logic.* Biotechnol Bioeng, 1999. **62**(6): p. 722-9.

82. Trelease, R.B., R.A. Henderson, and J.B. Park, *A qualitative process system for modeling NF-kappaB and AP-1 gene regulation in immune cell biology research.* Artificial Intelligence in Medicine, 1999. **17**(3): p. 303-21.

83. D'Haeseleer, P., et al., *Linear modeling of mRNA expression levels during CNS development and injury.* Pacific Symposium on Biocomputing, 1999. **163**(1): p. 41-52.

84. Mjolsness, E., D.H. Sharp, and J. Reinitz, *A connectionist model of development.* J Theor Biol, 1991. **152**(4): p. 429-53.

85. Varner, J. and D. Ramkrishna, *Metabolic engineering from a cybernetic perspective: aspartate family of amino acids.* Metab Eng, 1999. **1**(1): p. 88-116.

86. Voit, E.O. and T. Radivoyevitch, *Biochemical systems analysis of genome-wide expression data.* Bioinformatics, 2000. **16**(11): p. 1023-37.

87. McAdams, H.H. and A. Arkin, *Stochastic mechanisms in gene expression.* Proc Natl Acad Sci U S A, 1997. **94**(3): p. 814-9.

88. Srivastava, R., M.S. Peterson, and W.E. Bentley, *Stochastic kinetic analysis of the Escherichia coli stress circuit using sigma(32)-targeted antisense.* Biotechnol Bioeng, 2001. **75**(1): p. 120-9.

89. Arkin, A. and J. Ross, *Computational functions in biochemical reaction networks.* Biophys J, 1994. **67**(2): p. 560-78.

90. Swanson, C.A., A.P. Arkin, and J. Ross, *An endogenous calcium oscillator may control early embryonic division.* Proceedings of the National Academy of Sciences of the United States of America, 1997. **94**(4): p. 1194-9.

91. Rao, C.V., J.R. Kirby, and A.P. Arkin, *Diversity and Design in Bacterial Chemotaxis.* 2002: p. Submitted.

92. Doedel, E.J. *AUTO: A program for the automatic bifurcation analysis of autonomous systems*. in *Proceedings 10th Manitoba Conference on Numumerical Mathematics and Computation*. 1981. Univ. of Manitoba, Winnipeg, Canada.

93. Tonse, S.R., et al., *PRISM: Piecewise Reusable Implementation of Solution Mapping. An Economical Strategy for Chemical Kinetics.* Isr. J. Chem., 1999. **39**: p. 97106.